



# JOURNALISM HISTORY AND DIGITAL ARCHIVES

Edited by  
Henrik Bødker



# Journalism History and Digital Archives

This book showcases various ways in which digital archives allow for new approaches to journalism history. The chapters in this book were selected based on three overall objectives: 1) research that highlights specific concerns within journalism history through digital archives; 2) discussions of digital methodologies, as well as specific applications, that are accessible for journalism scholars with no prior experiences with such approaches; and 3) that journalism history and digital archives are connected in other ways than through specific methods, i.e., that the connection raises larger questions of historiography and power.

The contributions address cases and developments in Asia, South and North America and Europe; and range from long-range, big-data, machine-learning and topic modeling studies of journalistic characteristics and meta-journalistic discourses to critiques of archival practices and access in relation to gender, social movements and poverty.

The chapters in this book were originally published as a special issue of *Digital Journalism*.

**Henrik Bødker**, Ph.D, is Associate Professor at the Media and Journalism Studies Department at Aarhus University, Denmark. He has published on various intersections between popular culture and media, e.g. music and magazines. In addition to questions related to digital archives, he is currently working with how digital technologies and practices relate to changed patterns of circulation and new temporalities of journalism.



**Taylor & Francis**

Taylor & Francis Group

<http://taylorandfrancis.com>

# Journalism History and Digital Archives

*Edited by*  
**Henrik Bødker**

First published 2021  
by Routledge  
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge  
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

© 2021 Taylor & Francis

Chapter 3 © 2018 Marcel Broersma and Frank Harbers. Originally published as Open Access.  
Chapter 7 © 2018 Pernilla Severson. Originally published as Open Access.

With the exception of Chapters 3 and 7, no part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers. For details on the rights for Chapters 3 and 7, please see the chapters' Open Access footnotes.

Chapter 7 of this book is available for free in PDF format as Open Access from the individual product page at [www.routledge.com](http://www.routledge.com). It has been made available under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 license.

*Trademark notice:* Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

*Library of Congress Cataloging-in-Publication Data*  
A catalog record for this title has been requested

ISBN13: 978-0-367-56661-6

Typeset in Myriad Pro  
by codeMantra

#### **Publisher's Note**

The publisher accepts responsibility for any inconsistencies that may have arisen during the conversion of this book from journal articles to book chapters, namely the inclusion of journal terminology.

#### **Disclaimer**

Every effort has been made to contact copyright holders for their permission to reprint material in this book. The publishers would be grateful to hear from any copyright holder who is not here acknowledged and will undertake to rectify any errors or omissions in future editions of this book.

# Contents

<i>Citation Information</i>	vii
<i>Notes on Contributors</i>	ix
Introduction: Journalism history and digital archives <i>Henrik Bødker</i>	1
1 A Century of Journalism History as Challenge: Digital archives, sources, and methods <i>Thomas Birkner, Erik Koenen and Christian Schwarzenegger</i>	9
2 Excavating Concepts of Broadcasting: Developing a method of cultural research using digitized historical periodicals <i>James F. Hamilton</i>	24
3 Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency <i>Marcel Broersma and Frank Harbers</i>	38
4 In Search of America: Topic modelling nineteenth-century newspaper archives <i>Quintus Van Galen and Bob Nicholson</i>	53
5 Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism <i>Matthew S. Weber and Philip M. Napoli</i>	74
6 Saving Data Journalism: New strategies for archiving interactive, born-digital news <i>Meredith Broussard and Katherine Boss</i>	94
7 The Politics of Women's Digital Archives and Its Significance for the History of Journalism <i>Pernilla Severson</i>	110

8	Digital Archiving as Social Protest: <i>Dalit Camera</i> and the mobilization of India's "Untouchables" <i>Subin Paul and David O. Dowling</i>	127
9	Digital Archives as Subaltern Counter-Histories: Situating "Favela Tem Memoria" in the Rio de Janeiro media and political landscape <i>Stuart Davis</i>	143
10	@franklinfordbot: Remediating Franklin Ford <i>Juliette De Maeyer and Dominique Trudel</i>	158
	<i>Index</i>	177

# Citation Information

The chapters in this book were originally published in the *Digital Journalism*, volume 6, issue 9 (2018). When citing this material, please use the original page numbering for each article, as follows:

## Introduction

*Journalism history and digital archives*

Henrik Bødker

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1113–1120

## Chapter 1

*A Century of Journalism History as Challenge: Digital archives, sources, and methods*

Thomas Birkner, Erik Koenen and Christian Schwarzenegger

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1121–1135

## Chapter 2

*Excavating Concepts of Broadcasting: Developing a method of cultural research using digitized historical periodicals*

James F. Hamilton

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1136–1149

## Chapter 3

*Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency*

Marcel Broersma and Frank Harbers

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1150–1164

## Chapter 4

*In Search of America: Topic modelling nineteenth-century newspaper archives*

Quintus Van Galen and Bob Nicholson

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1165–1185

## Chapter 5

*Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism*

Matthew S. Weber and Philip M. Napoli

*Digital Journalism*, volume 6, issue 9 (2018) pp. 1186–1205



**Chapter 6**

*Saving Data Journalism: New strategies for archiving interactive, born-digital news*  
Meredith Broussard and Katherine Boss  
*Digital Journalism*, volume 6, issue 9 (2018) pp. 1206–1221

**Chapter 7**

*The Politics of Women's Digital Archives and Its Significance for the History of Journalism*  
Pernilla Severson  
*Digital Journalism*, volume 6, issue 9 (2018) pp. 1222–1238

**Chapter 8**

*Digital Archiving as Social Protest: Dalit Camera and the mobilization of India's "Untouchables"*  
Subin Paul and David O. Dowling  
*Digital Journalism*, volume 6, issue 9 (2018) pp. 1239–1254

**Chapter 9**

*Digital Archives as Subaltern Counter-Histories: Situating "Favela Tem Memória" in the Rio de Janeiro media and political landscape*  
Stuart Davis  
*Digital Journalism*, volume 6, issue 9 (2018) pp. 1255–1269

**Chapter 10**

*@franklinfordbot: Remediating Franklin Ford*  
Juliette De Maeyer and Dominique Trudel  
*Digital Journalism*, volume 6, issue 9 (2018) pp. 1270–1287

For any permission-related enquiries please visit:  
<http://www.tandfonline.com/page/help/permissions>

# Contributors

**Thomas Birkner** Department of Communication, University of Muenster, Germany.

**Henrik Bødker** Department of Media and Journalism Studies, School of Communication and Culture, Aarhus University, Denmark.

**Katherine Boss** University Libraries, New York University, USA.

**Marcel Broersma** Centre for Media and Journalism Studies, University of Groningen, The Netherlands.

**Meredith Broussard** Arthur L. Carter Journalism Institute, New York University, USA.

**Stuart Davis** Department of Communication Studies, Baruch College, City University of New York, USA.

**Juliette De Maeyer** Department of Communication, Université de Montréal, 90 avenue Vincent d'Indy, Montreal, Quebec, Canada.

**David O. Dowling** School of Journalism and Mass Communication, University of Iowa, USA.

**James F. Hamilton** Department of Entertainment and Media Studies, Grady College of Journalism and Mass Communication, University of Georgia, Athens, USA.

**Frank Harbers** Centre for Media and Journalism Studies, University of Groningen, The Netherlands.

**Erik Koenen** Centre for Media, Communication and Information Research, University of Bremen, Germany.

**Philip M. Napoli** Sanford School of Public Policy, Duke University, USA.

**Bob Nicholson** Department of English, History and Creative Writing, Edge Hill University, St Helens Road, Ormskirk, UK.

**Subin Paul** School of Journalism and Mass Communication, University of Iowa, USA.

**Christian Schwarzenegger** Department of Media, Knowledge and Communication, University of Augsburg, Germany.

**Pernilla Severson** Department of Media and Journalism, Faculty of Arts and Humanities, Linnaeus University, Kalmar, Sweden.

**Dominique Trudel** Department of Arts and Letters, Université du Québec à Chicoutimi, 555 boulevard de l'Université, Chicoutimi, Québec, Canada.

**Quintus Van Galen** Department of English, History and Creative Writing, Edge Hill University, St Helens Road, Ormskirk, UK.

**Matthew S. Weber** Hubbard School of Journalism and Mass Communication, University of Minnesota, USA.

# INTRODUCTION

## Journalism history and digital archives

Henrik Bødker

### **Journalism History *through* Digital Archives and across Fields and Institutions**

When putting together this special issue *Journalism History and Digital Archives* I had three related objectives and/or arguments: (a) that the selected articles should work with specific concerns within journalism history *through* digital archives; (b) that the discussion of digital methodologies, as well as specific applications, should be accessible for journalism scholars with no prior experiences with such approaches; and (c) that journalism history and digital archives are connected in other ways than through specific methods, i.e. that the connection raises larger questions of historiography and power. In the following I will briefly discuss these objectives as a way to both raise some general issues and to frame the selected articles; as a starting point for this discussion I will present a few reflections on what digital archives imply in this connection.

Traditionally, the term archive referred to collections of unpublished and unique documents or records (or artefacts) and not to published material, which were stored in libraries. Yet, journalism and other text-focused humanities scholars increasingly talk about archives rather than libraries, a widening that partly is related to the advent of the digital, but also to a “research tradition ... inspired by Derrida and Foucault” (Strandgaard Jensen 2017a, 70; my translation), which stressed the disciplining aspect of certain categories being deemed worthy as heritage. A more technology-related reason is the rise—since the mid-1990s—of web archiving, i.e. the “the act of collecting and preserving the online web and making it available” (Brügger 2018, 77). Yet, as most of the stored material is public, the web archive ought to be, Brügger jokingly asserts, called a “webrary” (78; emphasis in the original). A final reason for talking about “digital newspaper archives” (1) rather than libraries—as Steel (2014) does in the introduction to a special issue of *Media History* on “Digital Newspaper Archival Research”—is that approaching such repositories almost automatically concerns “developments and opportunities in the production, use and development of digital archives themselves” (Steel 2014, 1) as much as it concerns the stored content itself.

Given an interest in the broader institutional and political aspects of archives as well as an interest in journalism that is wider than newspapers the notion of a digital archive of journalism underlying this special issue can thus be described as rather simply as

digital archives that contain journalistic publications, productions or related content (in writing, still images, moving images or sound—or combinations thereof) stored and made available in digital form regardless of whether this is the result of digitization of whether the content was born as digital content.

I have here, as well be more obvious when discussing the different articles and the archives they focus on, notably not listed a feature of “large-scale digitized collections” that Gooding focus on in his recent *Historic Newspapers in the Digital Age* (2017, 3) namely that the content should be derived from “national library collections.” Such collections are of course very important but digital technologies have also made possible archives unrelated to more established institutions—as will be discussed below the heading “Reading Archives” (the third objective listed above).

The first objective listed above, e.g. the aim of collecting articles on journalism history and digital archives that arose from concerns related to journalism history *and* digital archives, is linked to the fact there already is a fair amount of theoretical work dealing with the more generic issues of digital methodologies and archives. The articles here are thus meant to highlight possibilities and pitfalls from the vantage point of journalism history, rather than focusing acutely on developing methods for monitoring and studying the contemporary and fluid landscape of digital journalism, a topic that was covered very well in the special issue “Rethinking Research Methods in an Age of Digital Journalism” of *Digital Journalism* (2016, 4(1) edited by Michael Karlsson and Helle Sjøvaag). Obviously, there are shared concerns of analytical methods and scale between studies of contemporary studies of digitally “live” journalism and those of digitised historical material; another issue that somehow breaches archival research and contemporary studies is that contemporary journalism has to be captured and stored in order to be studied and this may extend to the making of actual archives—as exemplified by Weber’s article in this issue. Yet, as Nicholson and Van Galen point out in this issue, there are indeed also important differences in terms of the structure and quality of data between digitised and digitally born material.

The second objective is aimed at distancing the work here from more detailed methodological discussions within the digital humanities and thus to highlight both challenges and opportunities in an accessible way for scholars who have not taken on computational approaches and who may otherwise have been put off by a steep learning curve standing between them and the potential outcomes of digital methods. In fact, and this is important, a number of the articles here precisely demonstrate the value of digital archives in ways that do not necessitate any prior engagement with more complex digital methods. But, utilising some of the possibilities opened up by digital archives does necessitate a reliance on computational processes that calls for new types of knowledge that scholars of (journalism) history traditionally have not had. This has—as in other fields—collectively contributed to the rise of digital humanities as scholars with various disciplinary interests collaborate to better understand methodologies at the intersections of “the digital” and “the humanities.” It is, however, arguably important that digital methodologies also are continuously developed and applied within specific fields, e.g. journalism studies and history, in addition to (or instead of) migrating to a separate and almost wholly methodology-focused field. This special issue highlights this need, as it reflects where specific legacies and considerations

unique to journalism's history texture the approaches which are useful, or even possible, when considering digital news archives.

The aim of having a special issue composed of articles situated within journalism studies and history raised a number of issues with regard to how new methodological possibilities can be written about—and for whom—and this draws out a further discussion among those working on digital archives which spans a number of fields: journalism studies, history, digital humanities, library and information studies and—to some extent—computer science. Specifically, this is an ongoing discussion among scholars working with digital archives whose scholarship and methods now cross from discipline to discipline. Given this, the peer reviews regularly split between one reviewer with a specific disciplinary background recommending “publish as it stands,” while another from a different academic field raised serious concerns as the articles were being read from rather different viewpoints. While a journalism studies scholar not particularly well grounded in computational methods would find an article a very informative illustration of how digital methods could be utilised within journalism history, another reviewer would find the application of computational tools wanting in terms of sophistication and nuance. The articles have thus been pushed to aim for a balance between introducing and applying digital methods in ways that are understandable to more conventional journalism scholars, while acknowledging the state of the art within the broader field of digital humanities—not always an easy balance to strike.

A related issue that emerged in this project are the cross-institutional interests in issues of archiving. While people working within archives focus on various (somewhat technical) issues of storing and making available different forms of content, journalism scholars may instead argue that such issues are better discussed in a journal of library studies and not within the field of journalism studies. Yet, knowledge of such processes is arguably of increasing importance for journalism studies, as in order for journalism scholars to utilise digital archives for collecting material and, not least, for making appropriate interpretations of this material, understanding the structures and accessibility of material is crucial. As many (most) digital archives are the products of complex agreements between public, individual, research and commercial interests, these necessitate scholars develop what Strandgaard Jensen (2017b) calls a “digital archival literacy,” i.e. an understanding of the processes underlying the digital material upon which you wish to base your research (a somewhat similar call is made by Birkner et al. in this issue). Linked to this is a need for journalism scholars to work with people involved in the storing, maintenance and dissemination of digitally stored journalistic products, as the public facing access, search functions, and data formats often prohibit more detailed analyses. Researchers are thus often in need of pulling out data in different formats and larger quantities and this naturally points towards collaboration with those working with libraries and repositories.

### **Reading Archival Content**

The relatively recent move of troves of archived documents as well as stores of published material into digital forms, alongside the increasing amount of digitally-born material, has rendered certain processes easier, yet these have also opened up opportunities that severely complicate what used to be a relatively individual and manual

process of “approaching an archive.” With regard to the products of journalism, the move towards the digital and digitisation has generally made both accessing and searching archives much easier but has also produced degrees of decontextualization and issues of scale as it becomes relatively easy to access vast amounts of material that simply cannot be processed manually. Obviously, even when this material was there in physical form the mechanics of access and analysis were of a nature that prohibited approaching a corpus as such.

Notions of scale and possible modes of analysis have caused an ongoing discussion of a shift from close reading of a discrete portion of an archive to distant reading of large volumes of material, and a possible subsequent shift from sampling to analysing a complete corpus. Such completeness, however, is often deceptive: Firstly, in the sense that not everything may have been stored and/or digitised and, secondly, since the quality of the OCR (optical character recognition) scanning may differ widely depending, on the one hand, on digitization at different periods of technological development and, on the other, on the type and quality of the scanned original, the corpus may become somewhat uneven. Yet, the possibilities of working with large amounts of data are real and alluring inasmuch as they may reveal patterns not likely to emerge from analysis of smaller samples. This, however, related to a (possible) move away from close reading and the subsequent decontextualization and lack of nuanced readings that can come with such a move.

Related to this is the question of how specific research questions relate to empirical inquiries now possible with digital archives. While traditional archival work through more focused sampling requires a relatively precise research question to narrow one’s focus, this is not necessarily the case when applying digital methods to larger amounts of data. As a number of the articles in this issue demonstrate, the ability to access and analyse vast amounts of content works in both directions, as this approach just as often raises new specific questions for research as much as it answers them. Thus, digital approaches to journalism history not only bring about questions of distant reading supplanting close readings, but also suggest what distant readings can reveal about new and interesting ways to approach specific texts and time periods in new ways. In a recent study of “fictional space” through computational methods, Tenen (2018) describes such an approach rather well when he writes that “the formal, computational methods ... occasion opportunities for close reading, and not just reading at scale. My methods are diagnostic, in that they identify areas of interest and unusual trends that require closer critical attention” (Tenen 2018, 120). The value of complex distant readings is thus arguably reliant on being applied against the background of deep contextual knowledge about the specific areas of (journalism) history in focus.

### **Reading Archives**

Another but somewhat related issue linked to the institutional settings of archives concerns the types and amount of material, and how this relates to interests and power. Here it is important to underline how the resources available in different settings vary both with regard to processes of the digitisation of stored material and the amount and diversity of what was stored in the first place. While at some level this may cause a “re-entrenchment of the traditional canon” and a “re-disappearance” of

marginalized content (Henderson 2017, 2) in the sense that the most popular material is digitised and made available, it is also important to remember that the dissemination of digital technologies has made new grass-roots and experimental archives possible. Thus, while the digital allows for new and illuminating historical trajectories of journalism's "core," i.e. developments of conventional, mainstream, male-dominated, national (political) news journalism based on archives at established research institutions, digital technologies have also allowed for the making of a broader range of emerging archives at the periphery of research institutions. When looking at the possibilities (and pitfalls) of digital archives of journalism it is thus important to include work focusing on smaller less "passive" repositories of specific material in relation to that deemed hegemonically relevant for established (national) archives. Such work can reflect on how such smaller archives make possible the crafting of voices at the periphery of an otherwise largely male, national and political journalism. Thus, while the scale of established archives requires specific distant-reading methods—e.g. topic modelling and machine learning—emergent and smaller archives rather call for broader approaches and discussions related to the politics of archiving in relation to gender and the subaltern.

### The Special Issue

Following the discussions briefly introduced above, the articles for this issue have been selected according to two intersecting dimensions: one running from archives at established institutions to new and experimental ones, and one running from relatively simple and traditional approaches to more sophisticated computational methods of journalism research. The first of these dimensions functions as the organising axis as the issue starts with work utilising established archives, going on to articles dealing with archiving and analysing specific journalistic content and then on to articles discussing alternative and more "active" archives. The issue starts with Thomas Birkner, Erik Koenen and Christian Schwarzenegger's "A Century of Journalism History as Challenge—Digital Archives, Sources, and Methods" in which they exemplify and discuss issues related to establishing an appropriate corpus for studying the development of the inverted pyramid structure in mainstream newspapers in Germany from 1914 to 2014. While pointing towards the potential benefits of a specifically defined longitudinal study, one of the important lessons of the article is its underlining of the problems of assembling a corpus from scattered and incomplete collections, how such data are "cleaned" and normalised, the importance of detailed historical knowledge and, not least, consequences for subsequent analyses.

The next article by James F. Hamilton is entitled "Excavating Concepts of Broadcasting; Developing a Method of Cultural Research Using Digitized Historical Periodicals" and takes its cue from Raymond William's book *Keywords* to utilise digital archives of newspapers and magazines in the US in order to trace the development of the notion of "broadcasting" from its original usage in agriculture to its updated reference to electronic dissemination of messages in the 1920s. As this study follows the term 'broadcasting' across a range of media and thus collections, Hamilton's article illustrates the value of combining a simple keyword search within a conceptually strong and historically grounded framework and as such complements more exploratory approaches employing elaborate digital methods. The two articles which follow each



experiment with and discuss the potential and pitfalls of two specific computational methods related to digital archives, and specifically address the question of scale. In "Exploring the Long-term Transformation of News. Machine Learning, Newspaper Archives and Journalism History," Marcel Broersma and Frank Harbers develop and discuss how machine learning may yield interesting results in a longitudinal study of the development of genres in Dutch journalism. While this article rightly argues for the importance of more longitudinal studies of the forms of journalism (as does the article by Birkner et al, which opens the issue), Broersma and Harbers' article provides great insight into challenges of developing and applying the method of machine learning. The central issue is here how you can train an algorithm to recognise and code specific genres based on various latent characteristics of journalistic texts.

The next article discusses and applies a different method, namely topic modelling. The central question posed by Bob Nicholson and Quintus Van Galen in "In Search of America: Topic Modelling Nineteenth-Century Newspaper Archives" is how, given the enormous amount of British newspaper texts containing the word "America," we may gain an overview of how America was journalistically embedded at the time, i.e. what themes (topics) were related to the country across the Atlantic? Their article does this by conducting four specific experiments, each of which applies topic modelling. This article and the one by Broersma and Harbers are both insightful and accessible discussions of the methods of topic modelling and machine learning. Further, while both articles argue for the potential of these methods, they also stress some common problems, not least the quality of OCR and issues related to the digital segmentation of articles.

Following these, the next two articles are focused on archiving specific types of contemporary journalistic content and discuss various aspects of web archiving. The first, "Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism" by Matthew S. Weber and Philip M. Napoli discusses a project in which they archived the websites of a range of local news outlets in the US in order to learn more about the development of digital journalism. Thus, while there indeed are interesting examples of analytical approaches, the bulk of the article is focused on the potentials and problems related to designing and archiving a corpus of websites for a specific study that addresses developments over time. As such, the article is a valuable contribution to what almost certainly will be an important approach within journalism studies. The next article, "Archiving, Data journalism, Web archiving, News applications, Born-digital news, Software preservation" by Meredith Broussard and Katherine Boss, is related in that it focuses on how to archive data journalism productions, i.e. interactive projects that allow users to explore a range of data. But, rather than doing experiments, this article focuses on understanding the digital infrastructure within which such productions are made in order to suggest possible paths for their archiving. As such, this article shines an important light on the complexity of such journalistic productions and, not least, the ways in which future researcher might, or might not, access them.

The final batch of articles shift focus away from what most often is understood as digital archives of journalism to raise important issues related to the politics of archiving. The four articles focus on, respectively, archives specifically tailored to women journalists and their work, what may be termed subaltern or grassroots archives in India

and Brazil and, lastly, a homemade archive made for illustrating and experimenting with issues of historiography. In "The Politics of Women's Digital Archives and its Significance for the History of Journalism," Pernilla Severson analyses two archives—one American and one Swedish—that give privileged access to women journalists. By looking at the digital affordances of these archives, Severson raises important questions about the institutional contours of female voices and power within the landscape of journalism history. By highlighting the gendered nature of digital archives this article is an important reminder about how various vectors of power produce the materials through which history is made.

This is also in focus in the next two articles, both of which look at archives deliberately made as correctives to the "history" recorded by mainstream journalism. In "Archiving as Social Protest: Dalit Camera and the Mobilization of India's 'Untouchables,'" Subin Paul and David Dowling very productively analyse important linkages between social movements and news archiving in what they call the "ensorious media climate" of India. The next article by Stuart Davis, "Digital Archives as Subaltern Counter-Histories: Situating 'Favela Tem Memoria' in the Rio de Janeiro Media and Political Landscape," very succinctly addresses similar issues, exploring linkages between a specific disadvantaged community and digital archiving in the favelas of Rio de Janeiro. Taken together, these two articles (joining Severson's) highlight how increasingly available digital technologies allow for the accumulation of local news as well other material can act as powerful correctives to the ways disadvantaged communities are journalistically portrayed (and subsequently archived) by the mainstream. These articles thus cast a light both back in time towards important lacunae in what has been stored as well as looking forward towards how digital technologies can allow for the recording of corrective views, which may or may not be incorporated into more established archival institutions. The final article closing this issue also focuses on a somewhat peripheral archive. In "@franklinfordbot: Remediating Franklin Ford," Juliette De Maeyer and Dominique Trudel use a homemade collection of material by and linked to Franklin Ford (1849–1918), an American journalist, entrepreneur and thinker who conceptualized circulations of media content that remain highly relevant today. De Maeyer and Trudel approach this archive in both orthodox and novel ways, including by designing a "bot" that tweets random excerpts from the archive. They consequently use the making of and the different approaches to the archive to raise important questions related to media history, remediation and digital archives. Taken together, the last four articles forcefully remind us that the linkages between journalism history and digital archives is not only made up of methodological concerns related to the (distant) reading of journalistic content or form but also to broader political and theoretical questions about establishing and using archives.

This short introduction and the brief run-through of the 10 articles in this issue hopefully gives credit to the breadth and complexity of the articles assembled here, not only in terms of how digital archives of journalism can be approached but also in relation to what constitutes a digital archive and, not least, the power relations involved in constructing and maintaining archives. Digital archives—in their various forms—will necessarily grow and become even more important objects and locations of study for understanding the history of journalism, its contemporary setting as well as its future trajectories. It is thus important that students and scholars of journalism are not intimidated by the complex

relations undergirding digital archives, or the ever-evolving and malleable complexities of access to and usage of such archives. It is my sincere hope that this collection not only gives an interesting snapshot of important work being done with digital archives but also—and even more importantly—that it helps initiate more journalism scholars into the analysis of digital archives and, as a possible consequence, introduce central aspects of digital methodologies in their teaching.

A final note should be addressed to those involved in making this special issue possible. While Bob Franklin encouraged my idea from the beginning (when he was still the editor of *Digital Journalism*) the reviewers of the proposal also saw its potential. The authors of the selected articles should certainly be praised for their patience with my ideas and concerns to which they reacted very productively. This could also be said for the reviewers including those who offered feedback on various versions of articles as they developed. And, finally, my thanks to Editor-in-Chief of *Digital Journalism*, Oscar Westlund, who engaged with the issues raised in a detailed manner, endured my more lenient approach to deadlines and, most importantly and consistently—when discussions threatened to veer off into adjacent fields—pulled us back into digital journalism studies.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

## REFERENCES

- Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, MA: The MIT Press.
- Gooding, Paul. 2017. *Historic Newspapers in the Digital Age*. Milton Park: Routledge.
- Henderson, Desirée. 2017. "Recovery and Modern Periodical Studies." *American Periodicals: A Journal of History & Criticism* 27 (1): 2–5.
- Steel, John. 2014. "Introduction." *Media History* 20 (1): 1–3.
- Strandgaard Jensen, Helle. 2017a. "Digitale Arkiver som medskabere i ny historieskrivning" [Digital Archives as co-creators in the writing of new history." In *Digitale Metoder [Digital Methods]*, edited by Kirsten Drotner and Sara Mosberg Iversen, 69–86. Copenhagen: Samfundslitteratur.
- Strandgaard Jensen, Helle. 2017b. "Storing Stuff, Structuring Stories. The power of digital archives in contemporary historiography." Keynote paper presented at the Danish DIGHUMLAB Conference, Copenhagen, November 7. A revised version of this is currently under review as an article for *American Historical Review*.
- Tenen, Dennis Yi. 2018. "Toward a Computational Archaeology of Fictional Space." *New Literary History* 49 (1): 119–147.

# A CENTURY OF JOURNALISM HISTORY AS CHALLENGE

## Digital archives, sources, and methods

**Thomas Birkner** , **Erik Koenen**  and **Christian Schwarzenegger** 

*Approaching journalism history through digital archives, digital sources, and digital methods is a demanding task for media historians, but also offers prospects. We explore some of the challenges and potential benefits in the light of a concrete research project that investigates journalism history in Germany from 1914 to 2014. The project focuses on the development of journalistic news storytelling following the inverted pyramid model. This paper mainly discusses the difficulties of assembling an adequate corpus. The German case is complicated, mainly because the country's violent history, with two World Wars and two dictatorships, has left several desiderata for historical journalism research. We subdivide a hundred years of journalism history into different phases, and for each of these we discuss different approaches with regard to the availability, accessibility, and usability of sources in digital form. We conclude that digital archives and digital sources open up new techniques for historical journalism research, including methods such as automated content analysis and text mining. Nevertheless, new technological and cultural environments of news pose genuinely new challenges and require new skills and literacies to cope with journalism history through digital archives.*

### Introduction

Journalism history research is hard work.<sup>1</sup> Reconstructing history via a trade that serves “for the day” and is not concerned with archiving sources for historians is difficult. In the early years of radio and TV, for instance, storage space was limited and expensive, which has produced enormous desiderata when researching journalism history. In Germany, World War II destroyed a tremendous number of sources and this adds to more general challenges. As journalism researchers, we know that the past is essential to understand the present and prepare for the future and we therefore have to cope with these challenges. Surrounded by gaps, we try to find new paths opened up by digitalization.

In times when much of the crisis of journalism (Blumler 2010; Brüggemann et al. 2016; Russial, Laufer, and Wasko 2015) is related to the internet era, digitalization in

general, and the World Wide Web in particular (for a very useful distinction between internet, the Web, and digitalization, see Brügger 2012a), we discuss the challenges but also the advantages of these developments for journalism historiography. Our interests are twofold: on the one hand, we focus on the digitalization of editorial content in analog forms, together with the infrastructure of the Web, which offers new possibilities for journalism historians. On the other hand, we discuss how online journalism has produced digitally born editorial content for more than two decades now. Collectively, we can term both types of news sources “digital reborn sources” (Brügger 2012b, 104), because the processes of archiving and making available sources to some degree change both types of sources.

In this paper, we reflect on the benefits but also the disadvantages (Bingham 2010) of writing journalism history through digital archives, sources, and methods, through a project that covers 100 years of German journalism, from the reinstallation of censorship at the beginning of World War I to the challenges of the twenty-first century on the World Wide Web. German journalism history is seldom told due to the country’s violent history and the sources destroyed during conflicts. This project investigates how the journalistic news format of the inverted pyramid has evolved over the decades of the last hundred years in Germany.

The aim of this paper is not to provide findings but rather to use the project to illustrate and exemplify some of the advantages and the disadvantages of digital sources. The paper reflects mainly on the problems of constructing a corpus of texts for the project. The existing digital archives are incredibly heterogeneous and cannot entirely cover the irrecoverable destruction of sources; yet, the archives can still help to discover entirely new methods for journalism history research. In this paper, we illustrate some of the new venues digital archives open for journalism historiography, but also reflect on the methodological and other research-related implications of the available and accessible sources and their peculiarities. We conclude that journalism history in particular, and communication history in general, will have to deal with digital archives and digital sources in the future, and that this is a challenge media historians need to be equipped for both intellectually and methodologically (Birkner and Schwarzenegger 2016; Koenen et al. 2018).

### **The Inverted Pyramid Model**

The research project we take as the point of departure for our methodological and source-critical reflections is conceptualized as a follow-up to a study researching the history of German journalism from the first newspaper in 1605 to the breakthrough of modern journalism in 1914, before the outbreak of World War I (Birkner 2012, 2016). The broader aim of this was an attempt to write journalism history as closely connected to social history.

Modern journalism emerged in the West at the beginning of the twentieth century as journalists developed new forms of writing news stories that “evolved in a culture that was being reshaped on all sides by advances and changes in science, technology, industrialization, education, religion and a host of other human activities” (Stensaas 2005, 49). That also applied to a new news format that flipped the ordinary way of storytelling upside down in the sense that modern news storytelling advanced

the chronology of events by starting with the essential fact(s). There are several explanations for the emergence of this model, e.g. "the telegraphic transmission of news may have provided a model of how news reporting might be more brief and interpretive" (Schudson 1982, 109). Moreover, more rational and faster news consumption by readers might have supported this new strategy (Pöttker 2005), which makes the argument for investigating journalism history as social history even stronger.

The inverted pyramid model is an integral part of research on the broader "form of news" (Barnhurst and Nerone 2001; Barnhurst 2012) and "objectivity as strategic ritual" (Tuchman 1971, 1978). For our project, we thus decided to focus on how the inverted pyramid model evolved over the decades of the twentieth century. This format became global as part of journalism as an "Anglo-American invention" (Chalaby 1996; see also Chapman 2005) along with the diffusion of the news paradigm (Høyer and Pöttker 2005; Schudson 1982, 2005). And, as Barnhurst and Nerone (2001, 21) explain, while the "capacity to change news designs had been available for quite some time, ... journalists considered the existing form of news fully functional," which stresses the interrelations of social and journalistic developments. Until now, however, we have no empirical evidence for how this form of news developed, and we consider the German case with its changing political systems throughout the twentieth century to be extremely interesting and we therefore focus on this form throughout the decades of the twentieth century characterized by different political systems (the Weimar Republic, National Socialism, and the Cold War in East and West Germany) and changing media environments from printed newspapers to radio, TV, internet, and the media convergence of the twenty-first century.

As we follow the inverted pyramid model through the decades, we address new opportunities offered through the digitalization of newspapers, including new techniques of distant reading and scrutinizing vast amounts of news content. The main focus of this paper is, however, on the difficulties of assembling an adequate corpus for our project, difficulties that raise more general questions for historical journalism research.

### Difficulties in Selecting a Sample

We see our study in line with Pöttker (2005), who saw "the inverted pyramid model established at the *New York Herald* in 1895 when nearly 30 percent of the articles with more than fifty words followed that model" (Birkner 2016, 159). German textbooks for journalism students did not include this form of news writing and continued to follow a chronology model. La Roche famously described how the *New York Times* and *Vossische Zeitung* from Berlin each reported the assassination of Franz Ferdinand and his wife in Sarajevo in 1914 (La Roche 2006). While in Germany, they died in the penultimate sentence of the news story, the *New York Times* wrote (La Roche 2006, 88): "Archduke Francis Ferdinand, heir to the throne of Austria-Hungary, and his wife, the Duchess of Hohenberg, were shot and killed by a Bosnian student here today." We were, however, somewhat skeptical about the broader implications of La Roche's example, which caused us to take a look at German news the day after the assassination. In 1914, German news was a very heterogeneous field, from small papers in the countryside up to the large new mass press in the big cities and we consequently coded a sample of eight newspapers selected according to geography, regional

diversity, typology of newspapers, and the size of their readership. However, sampling was also driven merely by accessibility. The center for German press research (Deutsche Presseforschung) in Bremen had most of the papers that we used for the first phase of the project. What we found—in opposition to the example given by La Roche—was that the inverted pyramid model was established in German journalism in 1914, especially in the mass press of the time (Birkner 2012, 2016). The percentages of articles in the sampled newspapers that followed the inverted pyramid thus ranged from 16 to 44 percent.

Barnhurst and Nerone (2001) characterize this style of news as relatively stable. The situation in German in the twentieth century is, however, not yet described. Besides the two World Wars, there are other influences on journalism history that might affect news storytelling. For example, the development of technology—the distribution of news using the telegraph, telephone, or satellite—is highly relevant. Other relevant aspects include the rise of the mass press, radio and TV and, finally, the internet, all of which have affected routine journalistic practices; as Weischenberg and Birkner noted (2015, 409), “different narrative structures and forms have emerged in print, broadcast, and interactive media.”

In order to investigate how the inverted pyramid has developed we divided the century from 1914 onwards into 10-year brackets as we detected events of social and political importance also relevant to media history in the years 1924, 1934, and so on. The goal is thus to analyze the news in different media every 10 years, using the following time divisions:

- 1924, shortly after the introduction of radio in Germany. How was radio news structured, and did this change news articles in the press?
- 1934, shortly after Hitler and the National Socialists seized power. Was their new regime already observable in the style of writing news stories?
- 1944, shortly after the Allies invaded Normandy. How were German news outlets alike in an already-destroyed country?
- 1954, shortly after the (West-) German team winning the football World Cup gave the new television medium a boost. Did this change news articles in the press and radio news broadcasts?
- 1964, shortly after the second German television station (ZDF) started. How was TV news structured then?
- 1974, when, again, the (West-) German team won the football World Cup, this time at home. How was the tournament perceived in (East-) German media?
- 1984, shortly after private broadcasting was established in Germany. Did this change news articles in the press and radio news broadcasts?
- 1994, shortly after the German unification and the process of Western media companies buying Eastern newspapers. How is news presented in a unified German media landscape?
- 2004, after the dot-com collapse. Has the acceleration of news distribution online changed the printed and broadcast news?
- 2014, after the distribution of news via social media immensely gained ground. Have the developments in online journalism turned the inverted pyramid model upside down?

Roughly, these 10 decades could be assembled in three phases that each offers unique opportunities, but especially challenges, for journalism historians. The first phase is the time of the two World Wars from 1914 to 1945, which due to the destruction of sources, especially during World War II, will here be referred to as *the age of scarcity*. The second phase covers the years of the division of Germany and Europe in the Cold War era from 1945 to 1990, here called *the age of eclecticism* due to the less than systematic ways in which news outlets have been archived and digitalized. The third phase covers the years from the beginning of digitalization until today, here called *the age of profusion* due to the enormous complexity and volume of available sources. However, before we address the difficulties and prospects of the respective phases for assembling an adequate corpus for our research, we reflect on the advantages and disadvantages of digitalization in general for journalism history.

### Journalism History in Digital Ways

Digitalization has created new archives for content across the media matrix. This development fundamentally changes the starting point of journalism research. The infrastructures and logic of digitization raise entirely new critical questions of the consequences of using digitally “reborn” sources for research (Brügger 2012b, 104) and for the possibility of new insights and research horizons. Media historians have just begun to discover and explore the potentials of digital repositories, sources, and methods for their work and to understand the required literacy and skills. Digital resources not only make it easier to find specific content, but also open them up in a machine-readable and thus entirely new way. Nicholson (2013, 64) describes how the “digital turn” is related to media history: “Sources are ‘remediated’ and not just reproduced. Though a digitized text may look familiar, it is not the same source; we are able to access, read, organize and analyze it in radical new ways.”

In this context, the shifts brought about by the digitalization of newspapers are often described as fundamental. Mussell (2017, 17–18) highlights three changes and advantages of newspaper retro-digitization: editorial possibilities of digital reproduction; automatic indexing of newspaper content via OCR (optical character recognition); and automatic indexing and structuring of newspaper collections via metadata. Yet, when looking more closely at the often euphorically described new sources and possibilities for historical newspaper research, it quickly becomes clear that the situation is very heterogeneous and characterized by striking national differences as well as “wild” digitizing, i.e. a range of non-standardized ways of digitizing journalistic content. There are, however, international pioneers such as Australia, Finland, Great Britain, the Netherlands, Austria, and the USA, which are pushing ahead with the digitalization of newspapers and have institutionalized collections that set “high standards in terms of indexing, presentation and use” (Seiderer 2010, 165). In contrast, “the digitization of newspapers in Germany is still in its infancy” (Seiderer 2010, 165)—a finding from 2010 that remains relevant. The goal is to develop common standards for digitalizing newspapers in Germany, including a national portal for digital newspapers (Blome 2018). Until now, however, the decentralized media structure of the postwar period have entailed many initiatives and digitalization projects with an explicitly local and regional focus, for example digiPress for Bavaria or the zeitpunkt.nrw project, which started at



the end of June 2018 and is about to digitalize historical newspapers from the region of North Rhine-Westphalia. Other archives have other specifications: the Friedrich Ebert Foundation of the German Social Democrats provides a collection of union and workers' papers (<https://www.fes.de/bibliothek/digitale-bibliothek/zeitschriften-digitalisierung/>), and the Hessian libraries and their information system (HeBIS) contain content from World War I, but only for Hessen—so again with a regional focus.

In practice, the focus as well as the design and the functions of newspaper portals frame research possibilities for digital journalism and press history research and must therefore be considered when we approach digital sources. Systematically, the technological evolution of digital press reading rooms allows us to distinguish between:

1. *flat portals*, providing more exhibitions than digital reading rooms, which are limited to the rudimentary possibilities of digitalization (simple browsing by date and title),
2. *deep but data-restrictive portals*, which offer improved possibilities for deep searching of digitalized newspaper content, but are still highly restrictive in the re-use of full-text data and search results, and
3. *open "virtual research environments,"* which offer a wide range of possibilities in using metadata, full-text, and search results.

Most of the German newspaper collections belong to the first and second types. For many institutions, preserving and protecting their collections appear as more important goals than making their collections widely usable as a source for research. Despite their heterogeneous nature, some of the digital newspaper portals in Germany go beyond image digitization and basic navigation via selection lists and the calendar function of the flat portals with keyword and full-text search options, and in some cases offer further functions for faceting and filtering search results. From a technical point of view, their greatest unfulfilled potential lies in the provision of search results according to the idea of openness—a necessary condition for historical journalism research in digital ways. Instead of an open approach to metadata and OCR, there is a tendency to keep the data at the core of the newspaper digitization under "lock and key" and to reduce the results of full-text searches to newspaper clippings represented in search lists that cannot be re-used digitally. As we will explain below, such technology critique is not only an essential part of a critique of the "digitalized newspaper" but is also relevant for the development of concrete research strategies for digital journalism research.

## **Difficulties in Different Decades**

### *The Age of Scarcity (1914–1945)*

The first phase, the years from the First to the Second World War, Churchill described as a Second Thirty Years War (Churchill 1949). Indeed, many media companies and their archives were destroyed during World War II. The dismal situation regarding the digitalization of German newspapers for this era severely impacts the task of approaching journalism through digital archives and sources. In this period the

status of newspaper digitalization, measured against the criteria of *availability*, *accessibility*, and *usability*, is very unsatisfactory and poses significant problems for our research as the sources in digital form are very fragmented, not representative, and limited by the digital resources themselves (Fickers 2013; Schwarzenegger 2012).

There is no common newspaper portal and no cooperative digitalization strategy in Germany and the many existing services and resources are therefore scattered and difficult to survey. Leading media and quality newspapers in regional or national contexts are over-represented as the structural pillars of the German press followed by the local newspapers, the popular mass press, and finally the party newspapers. Regarding time, the years 1933–1945 are often a blind spot in the newspaper portals. Even more severe is the fact that many portals draw strict limits regarding usability. Keyword searches and full-texts are not standard, and in most portals, digital research is limited to newspaper pages as simple images, which is a stark contrast to, for instance, the Europeana Newspapers Portal (<http://www.theeuropeanlibrary.org>), which is an integrative and much more productive resource for research.

The German corpus in this collection provides a total of 11 newspapers in full-text: three Berlin newspapers for the period 1914–1930 and eight Hamburg newspapers for the period 1914–1939. Due to the mix of mass newspapers and quality newspapers and the time frame, the Hamburg corpus is particularly interesting as a source of digital research. The biggest obstacle at the moment is the switched-off API and, therefore, the missing export function. At present, it is necessary to manually harvest the different source layers of the digital newspapers, especially the metadata and the full-text data. For automatic content analysis or text mining, the full-text needs to be pre-processed. The OCR-indexed full-texts are still very “dirty” and require manual “cleaning.” With the time intervals proposed in our project, it is nevertheless possible to develop an analytical strategy based on a relatively modest sample. For our project, we were able to find, for example, the *Hamburger Nachrichten* from 1924.

Another essential digital archive is the ZEitungsinFormationsYSTm (ZEFYS) at the Staatsbibliothek in Berlin. Again, this archive makes visible many of the problems of the decentralized German media landscape over the last centuries, including the destruction of newspapers, etc. For some of the newspapers, only a few issues are available, and only from particular years. However, we can find newspapers for every year of this period, for example the *Deutsche Allgemeine Zeitung* from Berlin. Unfortunately, this digital archive provides mostly digitalized page images and only in some projects fulltext options.

The complex media landscape and rudimentary remains of newspapers, which are digitally accessible as sources, do not only pose hindrances regarding local or regional representation and diversity, but also in terms of time. As stated above, the Europeana Newspapers Portal only provides the Berlin paper until 1930 and the Hamburg paper until 1939, which leaves us with the archive of ZEFYS and the newspapers they have for 1944, for example the *Leipziger Neueste Nachrichten* and the *Preußische Zeitung*. The latter reports on 31 March 1944:

The fact that the bombing of German cities and the civilian population is a conscious act of terrorism by the Roosevelt armies has been confirmed again in an impertinent manner by the ‘Philadelphia Record,’ known as a Jewish New Deal paper and the special mouthpiece of the USA-President.

Such a style of reporting may signal that the inverted pyramid model lost shape during National Socialism.

### *The Age of Eclecticism*

The second phase embraces the years between the end of the Second World War and the rise of digitalization. In this period, economic growth went hand in hand with a growing media sector, e.g. the spread of radio and later TV as well as the advertising industry. Those were good years for journalism—but they are not necessarily good years for journalism historians. That is why we call this phase *the age of eclecticism*. Both in radio and TV, storage media were extremely expensive and therefore frequently dubbed and re-used, which means that some news reports have left no trace. However, while radio and TV stations around the world, including in Germany, have started digitalizing what they did archive, they do not always make these collections accessible to the public, not even for research.

The situation for printed news seems better. As stated above, there are several initiatives to digitalize newspapers, including the second half of the twentieth century, and make them publicly available. One project of particular interest for our research is at the already-mentioned ZEFYS of the Staatsbibliothek in Berlin, that is especially concerned with the press of the GDR (German Democratic Republic), the socialist state in Eastern Germany that was part of the Soviet-dominated Warsaw Pact. Three main newspapers were digitalized, and their texts are fully accessible: *Neues Deutschland* (ND), from 23 April 1946 to 3 October 1990; *Berliner Zeitung* (BZ), from 21 May 1945 to 31 December 1993; and *Neue Zeit* (NZ), from 22 July 1945 to 5 July 1994.

For our analysis we look at 1954 and 1974—both years when, incidentally, West Germany won the football World Cup. The digital archives of Western and Eastern newspapers thus give us the opportunity to compare the diachronic development of news storytelling within both National Socialism and Western Democracy. When the *Berliner Zeitung* (from East Germany) (Figure 1) reported on the world cup it did so with a photograph and quite an unusual form of news storytelling, which seems a cliff-hanger or teaser for the content of the sports section: “A scene in front of the Hungarian goal in the big World Cup match Hungary-West Germany at the Wankdorf stadium in Bern, about which we report on the sports side.”

In 1974, the result was reported in a typically inverted pyramid model (Figure 2), which we can easily illustrate, as here the archive provides full-text access. The text started thus:

With a 2-1 victory of the FRG team over the representation of the Netherlands, the final match of the X. World Cup ended yesterday afternoon in front of 80,000 spectators at the Olympic Stadium in Munich.

Another very relevant digital archive is that of Frankfurt newspapers, notably the essential *Frankfurter Allgemeine Zeitung* (FAZ). This archive also represents all the different stages of the digitalization process and therefore serves as an example of the advantages and disadvantages regarding our research, providing articles in HTML format and article facsimiles in PDF format, and from 2001 onward, the entire newspaper in PDF format as well. For the victory in the World Cup finals in 1954 and 1974, we can see the change in newspaper design in the PDF format in the amount, and as well



FIGURE 1  
Title page of the Berliner Zeitung from 6 July 1954 (Staatsbibliothek zu Berlin – Preussischer Kulturbesitz: <http://zefys.staatsbibliothek-berlin.de/ddr-presse>).



FIGURE 2  
Source layers of the newspaper digital of the Berliner Zeitung from 8 July 1974: image and full-text (Staatsbibliothek zu Berlin – Preussischer Kulturbesitz: <http://zefys.staatsbibliothek-berlin.de/ddr-presse>).

as the style, of reporting in West German papers in the quality paper FAZ for these years.

The archive of the <https://www.faz-biblionet.de/> offers full newspaper pages in PDF format, which contain all graphics, photos, and images while ads, price tables,

weather news, event announcements, and publisher supplements are not provided. These articles are facsimiles in PDF format of full pages and entire issues offer potential, especially in the field of visual communication. In a historical perspective, they can help overcome deficits (Deacon 2007) of text-based journalism databases, the oft-stifled “productive tension between seeing and saying” (Maurantonio 2014). However, they are then also restricted in terms of automated analyses. In our context, it is primarily the HTML format that provides enormous possibilities for (automated) content analysis of the news storytelling.

### *The Age of Profusion*

The third phase of the project poses challenges that are entirely different from those of the first two. While we have until now discussed issues linked to working with originally analogue and then digitalized and thus digitally reborn sources, in this third phase we are dealing with journalism sources that have always been digital. Yet, although born digital this material must also be considered digitally reborn when approached historically; digitally born sources are through collection, preservation and displaying to some degree changed and must therefore be considered digitally reborn. Even if obtained from the newspaper itself, the preserved version is likely to differ somewhat from the version that originally appeared on the Web (e.g. adaption to new layouts, optimized for current browser versions, etc.). Accordingly, this phase is characterized by a totally new and confusing overload of sources—hence the term *the age of profusion*. The German team won the World Cup final again in 2014 and besides printed news and its HTML version we can also analyze online news. At first, online journalism was mostly a digital duplicate of, or addendum to, the analogous newspapers, and as such a “supplement and a complement to the dominant print and broadcast news media” (Scott 2005, 93). Increasingly, however, it developed its own logic and altered the face of production, dissemination, and reception of the news.

A significant aspect is here linked to how journalistic content is diversified and multiplied through the ways in which they reach audiences. The multiple links, teasers, and lead texts designed to lure an audience to an actual article have become essential parts of news formats. The digital news experience by users can be voluntarily customized or algorithmically personalized. This includes delivery and notification systems: push message news alerts or pushing news through mobile apps change the provision of news (Hermida 2010; Westlund 2013), and news is being optimized and formatted for mobile consumption.

In terms of research into the prevalence of the inverted pyramid in news reporting, the digital age challenges the specificity of the research unit. Is the actual news report the unit that we can rely on in our reconstruction, or do we need to consider the many potential pathways of how audiences could have approached it? When the purpose of the inverted pyramid is considered, the question of the initial contact with news content becomes more relevant. If the aggregators’ intermediaries leading to the actual report are considered part of the journalistic storytelling, other archives and archival strategies are necessary. If only the article as such is considered as a unit of comparison, tracing it can be easier, but critical reflections of the boundaries of articles as archived become increasingly important. Digitally born sources thus demand that

we reconsider the relationships between the historiography of the media and communication, and its materials and objects (Balbi 2011; Cowsls and Bright 2017; Weber 2017). In relation to this we also need to consider the question of who is collecting and who is granting access (e.g. legal regulations and copyrights; private, commercial, or public archiving). The lack of public archiving of web content and web journalism (in some contexts) constitutes only one among other unresolved problems.

The historical reconstruction of journalism in the digital age therefore demands that scholars carefully reflect on the limitations and potential of the source materials available to them. The analysis of the last few decades of journalism history is shaped by the dependence on (severely limited) digital archives, the preservation of all the visual and audio-visual dimensions of the news, and difficulties in capturing the complex entanglements and dynamics across various media which are characteristic of digital communication environments. Rather, the nature of data is dynamically linked to the process of its collection and is subject to change and inconsistencies through this process. Search engines and filtering processes generate data as custom-made products, with only very little insight available for how the data were gathered (Andrejevic, Hearn, and Kennedy 2015), and thus we face the problem of non-transparency regarding integrity, composition, and blind spots of the materials.

If we combine our findings for the criteria *availability*, *accessibility*, and *usability* with the three phases of *the age of scarcity*, *the age of eclecticism*, and *the age of profusion*, we come up with the following matrix:

	<i>Age of scarcity</i>	<i>Age of eclecticism</i>	<i>Age of profusion</i>
<i>Availability</i>	Rudimentary, locally diverse, temporally inconsistent	Better, some complete collections available, selective, disparate	High, abundance of potential sources, different formats and channels
<i>Accessibility</i>	Often only image digitization and basic navigation via selection lists and the calendar function, mostly no full-text	Possible, often restricted by private media companies	Generally high but diverse, expensive newspaper archives, changing social media platform policies; “lost internet” on mobile devices often inaccessible
<i>Usability</i>	Depends, often restricted to searching and reading, no standardized data format	Better, some media provide very good HTML versions	Depends, requires skills and literacy, various data formats

In the first phase—*the era of scarcity*—we are confronted with a historically based rudimentary preservation. Here this shortage is, on the one hand, being further exacerbated by digitization, which on the other hand facilitates access to archival records of news media outlets. The same applies for the second phase—*the era of eclecticism*—where eclectic selection from a much broader variety of possibly available titles is again narrowed down by digitization. Finally, the third phase—*the era of profusion*—with its wild growth of sources leaves us, as researchers, with the burden of selecting. That makes the task of assembling an adequate corpus very demanding in each of the three

phases, especially in order to investigate the development of the pyramid structure over all three phases.

### Concluding Discussion

This paper has discussed advantages and challenges of digital archives, sources, and methods for journalism history; special attention was given to the construction of an adequate corpus. We should remember, however, that “Historians deal with incomplete, biased data; but so do scientists” (Nord 1989, 311). Yet, we tentatively conclude that despite all the problems mentioned journalism history research in general and our project in particular will (potentially) benefit enormously from digital archives. If full-text is accessible, comparison of news formats becomes far easier, and in total we can then work with much more data. The potential but also risks of big data have been discussed, for example by Boyd and Crawford (2012), Mahrt and Scharkow (2013), and Lazer and Radford (2017). Yet, one can hope that internationally oriented, transnational, interdisciplinary journalism history will become, at least in parts, a computational social science which initially “needs to be the work of teams of social and computer scientists” (Lazer et al. 2009).

Such computational approaches continuously need to be complemented by qualitative analysis, close reading, and special in-depth scrutiny of particular sources. Digital methods also allow the development of analytic frames for rather small batches of sources, which then can be scaled toward bigger samples. We have explained how the possibilities and particular limitations brought about by digital archives and digital-born sources change specific features of sources and approaches to journalism history regarding the *availability*, *accessibility*, and *usability* of materials. But questions regarding the validity, reliability (Deacon 2007), and consistency of sources/data are not resolved by digital means, nor merely transferred into new technological and cultural environments of news. They pose genuinely new challenges and require new skills and literacies to be coped with.

Almost unlimited text corpora offer new possibilities for automated content analysis, which may add a new dimension to the analysis of the evolution of the inverted pyramid model. This format has been criticized in the same way as the information-centered journalism it represents, namely that it may easily survive dictatorship as a “strategic ritual” (Tuchman 1971, 1978) for objectivity. However, a complex phenomenon such as objectivity is not sufficiently described by the inverted pyramid model itself. In order to get at this broader question we need a combination of detailed historical knowledge, standardized research methods of social science, including automated content analysis and methods of historical research.

### NOTE

1. The authors would like to thank three anonymous reviewers and the editors for their helpful comments.

### ORCID

Thomas Birkner <http://orcid.org/0000-0002-0818-3062>

Erik Koenen <http://orcid.org/0000-0001-9824-385X>

Christian Schwarzenegger <http://orcid.org/0000-0003-1118-9948>

## REFERENCES

- Andrejevic, Mark, Alison Hearn, and Helen Kennedy. 2015. "Cultural Studies of Data Mining: Introduction." *European Journal of Cultural Studies* 18 (4–5): 379–394.
- Balbi, Gabriele. 2011. "Doing Media History in 2050." *Westminster Papers in Communication and Culture* 8 (2): 154–177.
- Barnhurst, Kevin G. 2012. "The Form of Online News in the Mainstream US Press, 2001–2010." *Journalism Studies* 13 (5–6): 791–800.
- Barnhurst, Kevin G., and John Nerone. 2001. *The Form of News: A History*. New York, NY: Guilford.
- Bingham, Adrian. 2010. "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." *Twentieth Century British History* 21 (2): 225–231.
- Birkner, Thomas, and Christian Schwarzenegger. 2016. "Konjunkturen, Kontexte, Kontinuitäten. Eine Programmatik für die Kommunikationsgeschichte im digitalen Zeitalter." *Medien & Zeit* 31 (3): 5–16.
- Birkner, Thomas. 2012. *Das Selbstgespräch der Zeit. Die Geschichte des Journalismus in Deutschland 1605–1914*. Köln: Halem.
- Birkner, Thomas. 2016. "Journalism 1914. The Birth of Modern Journalism in Germany a Century Ago." *Journalism History* 42 (3): 153–163.
- Blome, Astrid. 2018. *Zeitungen*. In *Clio Guide – Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften*, edited by Laura Busse, Wilfried Enderle, Rüdiger Hohls, Thomas Meyer, Jens Prellwitz, and Annette Schuhmann, [B.6] 1–36. Berlin: Clio-online and Humboldt-Universität zu Berlin.
- Blumler, Jay G. 2010. "Foreword. The Two-Legged Crisis of Journalism." *Journalism Studies* 11 (4): 439–441.
- Boyd, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–679.
- Brüggemann, Michael, Edda Humprecht, Rasmus Kleis Nielsen, Kari Karppinen, Alessio Cornia, and Frank Esser. 2016. "Framing the Newspaper Crisis. How Debates on the State of the Press Are Shaped in Finland, France, Germany, Italy, United Kingdom and United States." *Journalism Studies* 17 (5): 533–551.
- Brügger, Niels. 2012a. "Web History and the Web as a Historical Source." *Zeithistorische Forschungen/Studies in Contemporary History* 9 (2): 316–325.
- Brügger, Niels. 2012b. "When the Present Web Is Later the Past: Web Historiography, Digital History, and Internet Studies." *Historical Social Research* 37 (4): 102–117.
- Chalaby, Jean K. 1996. "Journalism as an Anglo-American Invention: A Comparison of the Development of French and Anglo-American Journalism, 1830s–1920s." *European Journal of Communication* 9 (11): 303–326.
- Chapman, Jane. 2005. *Comparative Media History*. Cambridge: Polity.
- Churchill, Winston S. 1949. *The Second World War. Volume I: The Gathering Storm*, London: Cassell & Co.
- Cowls, Josh, and Jonathan Bright. 2017. "International Hyperlinks in Online News Media." In *The Web as History: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 101–116. London: UCL Press.



- Deacon, David. 2007. "Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis." *European Journal of Communication* 22 (1): 5–25.
- Fickers, Andreas. 2013. "Veins Filled with the Diluted Sap of Rationality. A Critical Reply to Rens Bod." *BMGN – Low Countries Historical Review* 128 (4): 155–163.
- Hermida, Alfred. 2010. "Twittering the News: The Emergence of Ambient Journalism." *Journalism Practice* 4 (3): 297–308.
- Høyer, Svennik, and Horst Pöttker. 2005. *Diffusion of the News Paradigm 1850–2000*. Gothenburg, Sweden: Nordicom.
- Koenen, Erik, Christian Schwarzenegger, Lisa Bolz, Peter Gentzel, Leif Kramp, Christian Pentzold, and Christina Sanko. 2018. "Historische Kommunikations- und Medienforschung im digitalen Zeitalter. Ein Kollektivbeitrag der 'Initiative Kommunikationsgeschichte' digitalisieren zu Konturen, Problemen und Potentialen kommunikations- und medienhistorischer Forschung in digitalen Kontexten." *Medien & Zeit* 33 (2): 4–19.
- La Roche, Walther von. 2006. *Einführung in den praktischen Journalismus: mit genauer Beschreibung aller Ausbildungswege; Deutschland, Österreich, Schweiz*. Berlin: Econ.
- Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Davon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323: 721–723.
- Lazer, David, and Jason Radford. 2017. "Data Ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43, 19–39.
- Mahrt, Merja, and Michael Scharkow. 2013. "The Value of Big Data in Digital Media Research." *Journal of Broadcasting & Electronic Media* 57 (1): 20–33.
- Maurantonio, Nicole. 2014. "Archiving the Visual: The Promises and Pitfalls of Digital Newspapers." *Media History* 20 (1): 88–102.
- Mussell, James. 2017. "Beyond the 'Great Index.' Digital Resources and Actual Copies." In *Journalism and the Periodical Press in Nineteenth-Century Britain*, edited by Joanne Shattock, 17–30. Cambridge: Cambridge University Press.
- Nicholson, Bob. 2013. "The Digital Turn. Exploring the Methodological Possibilities of Digital Newspaper Archives." *Media History* 19 (1): 59–73.
- Nord, David P. 1989. "The Nature of Historical Research." In *Research Methods in Mass Communication*, edited by Guido H. Stempel and Bruce H. Westley, 290–315. Eaglewood Cliffs, NJ: Prentice Hall.
- Pöttker, Horst. 2005. "The News Pyramid and Its Origins from the American Journalism in the 19th Century: A Professional Approach and an Empirical Inquiry." In *Diffusion of the News Paradigm 1850–2000*, edited by Svennik Høyer and Horst Pöttker, 51–64. Gothenburg, Sweden: Nordicom.
- Russial, John, Peter Laufer, and Jane Wasko. 2015. "Journalism in Crisis?" *Javnost-The Public* 22 (4): 299–312.
- Schudson, Michael. 1982. "The Politics of Narrative Form: The Emergence of News Conventions in Print and Television." *Daedalus* 111 (4): 97–112.
- Schudson, Michael. 2005. "The Emergence of the Objectivity Norm in American Journalism." In *Diffusion of the News Paradigm 1850–2000*, edited by Svennik Høyer and Horst, 19–35. Göteborg: Nordicom.

- Schwarzenegger, Christian. 2012. "Exploring Digital Yesterdays – Reflections on New Media and the Future of Communication History." *Historical Social Research* 37 (4): 118–133.
- Scott, Ben. 2005. "A Contemporary History of Digital Journalism." *Television and New Media* 6 (1): 89–126.
- Seiderer, Birgit. 2010. "Die Digitalisierung von Zeitungen im deutschsprachigen Raum – ein Zustandsbericht." *Zeitschrift für Bibliothekswesen und Bibliographie* 57 (3–4): 165–171.
- Stensaas, Harlan S. 2005. "The Rise of the News Paradigm." In *Diffusion of the News Paradigm 1850-2000*, edited by Svennik Høyer and Horst Pöttker, 37–49. Göteborg: Nordicom.
- Tuchman, Gaye. 1971. "Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity." *American Journal of Sociology* 77 (4): 660–679.
- Tuchman, Gaye. 1978. *Making News: A Study in the Construction of Reality*. New York, NY: Free Press.
- Weber, Matthew S. 2017. "The Tumultuous History of News on the Web." In *The Web as History: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 83–100. London: UCL Press.
- Weischenberg, Siegfried, and Thomas Birkner. 2015. "News Story." In *The Concise International Encyclopedia of Communication*, edited by Wolfgang Donsbach, 408–409. Oxford: Wiley-Blackwell.
- Westlund, Oscar. 2013. "Mobile News: A Review and Model of Journalism in an Age of Mobile Media." *Digital Journalism* 1 (1): 6–26.

# EXCAVATING CONCEPTS OF BROADCASTING

## Developing a method of cultural research using digitized historical periodicals

**James F. Hamilton**

*This article demonstrates a method of cultural–historical analysis in which digital archives of magazines and newspapers play a central role. It addresses the formation of broadcasting as a particular complex of social relationships and, further, that can be addressed by investigating changes in the concept of broadcasting itself. Primary sources include digital archives of newspapers and magazines, supplemented by appropriate additional primary and secondary sources. The key question is how a late-1700s agricultural conception of broadcasting as the hand-spreading of seeds in a field emerged in the 1920s radio era and beyond as the indiscriminate electronic distribution of messages. The study documents a gradual, sedimented, radically historical process of mystification and naturalization, in which broadcasting is transmogrified from a human practice of agriculture (an intrinsically human activity) into the electromagnetic basis of modern media systems (a phenomenon of nature and thus by definition outside human control and beyond question or change). The article concludes with reflections about the research process described here, as well as about the value more generally of critical commentary regarding historical methods.*

This article theorizes, demonstrates and critiques a method of cultural–historical analysis used in a study regarding broadcasting, and in which digital archives of magazines and newspapers play a central role (Hamilton 2007). It addresses the foundational mass-communication topic of the formation of broadcasting. However, unlike conventional institutional histories, which trace its beginnings primarily through the development of technologies and/or the founding of companies (Barnouw 1966–70; Aitken 2014), this article views broadcasting more broadly as the institution of a complex of social relationships reproduced through cultural and material means and, further, one that can be productively addressed via an inquiry into the evolution of the concept itself (Scannell 1991; Peters 1999).

The insights gained by regarding media and communication generally through such a perspective and approach can be seen in a number of studies (Czitrom 1982; Covert 1984; Mander 1984; Marvin 1987; Spinelli 2000; McCauley 2002; Gitelman and

Pingree 2003; Thorburn and Jenkins 2003). But room exists to add to this oeuvre. As will be discussed, an even more thoroughly materialist approach to investigating media institutions and practices as a complex of social relationships can be inspired by the work of cultural theorist Raymond Williams and the method pioneered in his influential study *Keywords: A Vocabulary of Culture and Society* (Williams 1976).

Digital archives of historical newspapers and magazines—a resource unavailable to Williams—provide an as-yet largely untried resource for extending the reach and comprehensiveness of such a method due to their being full-text searchable. By contrast, and despite its innovation, his project was limited in scale and scope due to relying on scattered examples of usages gleaned from his own reading and from the unabridged Oxford English Dictionary (Williams 1976, 17).

After noting the instigation of the study in present-day concerns about concentration of ownership and hypercommercialization of media, and briefly outlining and critiquing a pilot study, the article describes the development and use of a revised and more extensive method used to investigate the emergence of broadcasting. The key question is how a late-1700s agricultural conception of broadcasting as the hand-spreading of seeds in a field coincides or intersects with the 1920s radio-era conception and beyond as the indiscriminate electronic distribution of messages, and with what implications for broadcasting as a democratic media institution and practice. Key sources included digitized historical periodicals and secondary sources. The article concludes with reflections about the research process described here, as well as about the value more generally of critical commentary regarding historical methods.

## Bases

The practice of media history should itself always be understood historically, and this case is no different. While granting the painful obviousness of the claim, the present study was instigated by the recognition of the steadily shrinking accountability of major media institutions to the societies and publics in and through which they operate, although this is not without its contradictions and potentials for change (Hamilton 2008). Despite the technical (at least) potential for more democratic media institutions, this goal has still has yet to be realized, despite hopes pinned, often and for example, on social media (Myers and Hamilton 2014).

The resulting key, general question became: Why? If the technical potential exists for democratic systems, what might steer and however-provisionally fix media technologies, institutions and uses into quite undemocratic forms?

Of course, there is no single, isolated cause or reason why, and this recognition led me to first clarify my theoretical perspective. To escape simplistic cause-and-effect thinking (which by necessity posits the “cause” as an ahistorical naturally-occurring entity), I viewed the problem as a cultural–historical one in the broadest terms. Media institutions and practices are not hard-wired to be a certain way. Instead, they should be recognized as radically historical, and thus emerging, forming and becoming realized through very specific historical conditions that they themselves help produce as well. Conceptions, practices, technical devices and institutions need to be regarded as mutually constituted and reproduced as a whole provisional and contradictory complex (Williams 1989, 173; Hardt and Brennen 1993; Slack and Wise 2002).

Upon solidifying the instigation, key question and the perspective through which I chose to address the problem, what was needed was a concrete and specific focus for the study. Broadcasting is tailor-made, given its status as the quintessential electronic mass medium as well as its promising technical potential but severely compromised implementation. Selected commentary separated by decades provides provocative justification for doing so. In one example, users found it strange and difficult to themselves talk using broadcasting, even when they had a non-industrial means available for doing so. Reflecting upon an incident in the 1990s, Hiken noted the difficulty of conceiving what to do with a microradio transmitter, due to being so used to listening “passively to what was broadcast over the airwaves,” that “only when ‘permitted’ or ‘invited’ to speak over an approved station” would he feel empowered or able to speak (Hiken 1999, x). In a second, earlier example from 1930, the accepted one-way use of broadcasting was recognized as only one of many potentials. In it, Brecht put the point well by observing that “the technique for all such projects has still to be developed; but it will be directed towards the prime task of ensuring that the public is not only taught but must also itself teach” (Brecht 1983, 171).

Now that a key question, theoretical perspective and focus for the study was established, the method for conducting it had to be specified as well. And it is here that the path-breaking work of cultural theorist Raymond Williams became relevant for the kind of cultural–historical excavation I was to attempt of the emergence of the modern conception of broadcasting. Although unaddressed by Williams, I suspected early on that “broadcasting” might valuably be investigated as a keyword in the sense he developed.

Williams’s book *Keywords* suggests the methodological usefulness of an inquiry into not simply inert and abstract meanings, but into a living, historical vocabulary and its corresponding key social formations. In his own extensive reading of accounts of culture by United Kingdom figures in 18- to early-20th centuries, Williams began to take notes on the different usages of certain recurring and central words, such as democracy, determine (as in determination), naturalism and, famously, culture (Williams 1976, 12). By understanding culture as a material practice—as something people do instead of as ideas solely in people’s heads—he viewed language as one such material practice (Williams 1977). Seen in this way, he realized that what he was collecting in his notes was not simply a set of meanings, but a record of deeply cultural and historical activities that were “explicit but as often implicit connections which people were making, in what seemed to me, again and again, particular formations of meaning—ways not only of discussing but of seeing many of our central experiences” (Williams 1976, 13).

In his entries for each keyword, he also briefly contextualized particular formations of meaning within their corresponding particular historical contexts in order to show the mutual determination of usages and their contexts. Thus, these keywords do not simply indicate what some individuals thought. When recognized as social and material usages in specific contexts, these words bind together other meanings, corresponding activities and their manifestations, which can emerge as institutions and systems. Of course, many manifestations cannot be captured in an historical record, seeing as they are immanent only in direct, lived experience. But at least an aspect of them can be available for critical–historical analysis in the form of writing and other traces of

corresponding social formations at the time. It was this integrative analysis of historical usages contextualized in their particular social formation that I planned to attempt.

To help prepare for my effort to extend Williams's approach in a cultural-historical analysis of broadcasting, I spent some time considering an earlier short trial run of mine that focused on the term "politically correct" as it had ascended in the U.S. by the late 1980s and early 1990s. In it, I executed a full-text search of the phrase "politically correct" in the Lexis-Nexis Academic database (now known as Nexis Uni), limiting results to 1980–1990. I then read enough occurrences in context to detect patterns validated by the repetition and resulting redundancy of distinct inflections. The resulting short essay described these different inflections, then ended by drawing some overall conclusions (Hamilton 1991).

While this approach yielded some insights, upon reflection I recognized a number of problems. For one, it lacked a connective analysis of the social formation of these meanings, offering instead at best a brief, introductory nod to the "current times." It also neglected to consider possible patterns of relationship between particular inflections and the narrative form of the accounts in which certain inflections appeared. Finally, by conceiving of context (society) and text (ideas/beliefs) as distinct, it reproduced the debilitating dualism of conventional, idealist approaches to cultural analysis in which "ideas" are seen to be either causes or symptoms of—and in either case wholly separate from—so-called "real-world" developments, instead of as a radically historical material practice of "human sociality" (Williams 1977, 165).

## Endpoints

To launch the study of broadcasting as a keyword, I addressed these drawbacks by retooling my method. Importantly, upon the availability on many university campuses by the 2000s of digital, full-text-searchable archives of historical periodicals (rather than just current publications, as was the case in the 1990s), resources now existed to extend the approach to cultural-historical analysis pioneered in *Keywords*.

Forearmed with my trial run and critique, my first step was to establish chronological beginning and end points. To do this, I consulted OED (the unabridged Oxford English Dictionary) entries along with relevant secondary works. Upon searching for "broadcast" (finding through experimentation that the verb form had a greater chronological range of entries than other forms), the earliest entries established evidence of its usage in England in 1807 as "to scatter (seed, etc.) abroad with the hand" (OED Online 2017). A robust secondary literature in agricultural-social history documented its complex economic role and class position at that time (Collins 1989; Davison et al. 1992; Mingay 1994; Mingay 2002). By this time, gentry farmers sought to phase out broadcasting in favor of mechanical seed drills in order to cheapen and rationalize production, as well as to reduce their reliance on peasant laborers. Indeed, broadcasting produced a dependency on the local peasant workforce, which might withhold its labor to force changes in employment conditions.

The secondary literature also established the significance of how these calls were circulated, with the main means being English farming books. In the words of one historian, these farming books were "the classics, part of every gentleman's education of the day" (Fussell 1947, 1). They consisted of first-hand descriptive accounts often self-published

of gentry farmers' experimentation with different techniques and methods, complete with personally-gathered statistical records as evidence (an example is Anstruther 1796). Due at least to their exclusivity (requiring money to purchase and literacy to appreciate), they helped produce a self-aware class consciousness commensurate with fitting into the mercantile and later the early capitalist economy.

In contrast to gentry farmers, peasant laborers had their own rationale for supporting broadcasting instead. Far from being an irrational response as gentry would have it, broadcasting preserved a means of rural livelihood. Not only did mechanical seed drills often break, thus delaying if not preventing the timely sowing of fields, they made much agricultural labor superfluous, as a few men using an oxen-drawn seed drill could plant areas that took scores of people who were broadcasting seed.

To locate the end point, the OED entry also established not surprisingly the 1920s as the emergence of the modern meaning of broadcasting as applied to radio networks, "to disseminate (a message, news, a musical or dramatic performance, or any audible or visible matter) from a radio or television transmitting station to the receiving sets of listeners and viewers" (OED Online 2017). One of the textual examples offered in the OED for this meaning struck me immediately as provocative because it focused on a type of social relationship. By noting that broadcasting in the early 1920s referred to "sending out messages without receiving replies," it suggested a wholly undemocratic form and relationship through its top-down, one-way conception (OED Online 2017). A robust secondary literature informed this sense of broadcasting as well as its formative context of industrialization and commercialization (or resistance to it) (Barnouw 1966–70; Briggs 1985; Hilmes 1990; Hilliard and Keith 2001).

### **Filling in**

However much value a combination of OED examples and secondary literature had in establishing the endpoints of my inquiry, what had yet to be established was how the agricultural sense of broadcasting came to take on the mass-media sense. It was here that digital historical archives of periodicals became crucial.

### *Delineating boundaries of the search*

Awareness of secondary literature about the rise of mass media in the 19th Century U.S. (sandwiched as it is between the 18th- and 20th-Century endpoints already identified) helped delineate the boundaries of the search. Studies that loomed particularly large (and that I was already familiar with—emphasizing the value of being aware of media historiography) was work regarding religious publishing in the early 19th-Century United States (Nord 1984, 2004; Brown 2004). Historians noted how evangelical publishers' practice embodied many presumptions of what later would be called broadcasting. Bible-tract societies pioneered mass means of financing, distribution and manufacture at a time earlier than the beginnings of the commercial penny press, which is commonly regarded as the initial mass, commercial medium. Nord concludes that they sought to "place in the hands of every city resident at least one tract, the same tract, each month," thus exemplifying "the ideology of the modern mass media:

to have everyone reading and talking about the same thing at the same time" (Nord 1984, 20).

These historical studies note in passing that religious-tract societies of the day did not call what they did at the time "broadcasting." Or did they? I suspected that the media-centric sense of broadcasting could not have come out of thin air. Given the similarity between the intentions and organization of evangelical publishing and broadcasting, I suspected that there had to be some evidence of its mention at this time. So, instead of taking it for granted, I thought it best to confirm. In order to cast the widest net possible, and after some trial runs using other online databases, I chose the database Accessible Archives (<https://www.accessible.com>), which had a robust set of magazines and journals full-text searchable in 1800–1875.

After executing a full-text search of "broadcast\*" (the asterisk to include variant spellings/forms), and contrary to points made by historians of evangelical publishing in the 19th Century, the term "broadcast" appeared in 163 items in a range of publications, and in the context of media dissemination. Most of these publications were in the abolitionist press, thus articulating social reform through a civic-religious lens. To preserve bibliographic information for later citation while also enabling chronological sorting (the hit list was not initially in chronological order), I downloaded all 163 records, then formatted them to fit a database composed of four fields for each record: its unique identification number, date of publication (a numerical value so as to enable chronological sorting), full title and publication of item, and the passage that contained the broadcast-related mention.

### *Fashioning an analytic strategy of interpretation*

The easy part was now done. I had the body of materials, and I had at least established that writers used the term "broadcast" in reference to pre-electronic media. The methodological problem that remained was how to move past the descriptive level to the analytic. Similar to my early effort regarding "politically correct," a descriptive method would have consisted of grouping then characterizing the range of ways "broadcast" was used in relation to media. However, doing this would replicate the shortcomings I had recognized earlier.

To move past these limitations, in light of my theoretical perspective and research already completed, and after repeatedly poring over in chronological order the entries I had found, I slowly developed a way of interpreting the mentions of "broadcasting" that addressed its mystification. The term mystification is relevant to a number of academic fields. Its specific relevance to this project concerns its place in critical traditions of cultural-ideological analysis (Williams 1977, 55–71). Put simply and deriving in good part from Marx's reflections on the commodity form (Marx 1967), mystification in this sense refers to the obscuring of the human origins, decisions and actions behind what in actuality is a human-produced process or institution, to the point where the process or institution seems to be virtually a fact of nature.

Recognizing the relevance of mystification helped me formulate a tentative response to my initial question regarding why, given the potential in its immense technological and organizational resources, broadcasting emerged and largely remains an undemocratic communications system. A mystified institution is one that asserts



that humans have no control over it and thus cannot change. If human conceptions, decisions and actions come to be regarded as products of nature, what has occurred is a mystification of (in this case) broadcasting's human basis, which in turn leads to a naturalization of its legitimacy, thus placing it beyond question, critique, challenge or change.

Mystification thus presents a much more difficult situation for media-reform efforts than lack of information (which can be easily be addressed by simply supplying it) or the application of some kind of external propaganda (which can be dealt with by simply removing it) (Williams 1980, 37). Mystification as a resonant and ostensibly commonsense way of instituting and practicing a large-scale communications system such as broadcasting helped account both for its cultural basis and its resistance to other, more democratic forms and uses.

Once I had worked out the direction my interpretation needed to take, I focused it further by formulating a key question I posed when reading yet again each of my 163 items individually, sequentially and in relation to the entire set. Does a particular usage of broadcast explicitly recognize that people do it? Or is broadcasting talked about as a natural force unrelated to human intention and activity?

### *Interpretation*

After returning to my found records, this heuristic allowed me to construct and describe in some detail the formation of a sense of broadcasting alongside—not in place of—the agricultural as a mystified, naturalized process of one-way message dissemination. When viewing this dynamic against its corresponding social formation, I saw the class indexing performed by broadcasting beginning to complicate in comparison to its agricultural usage and context. Broadcasting in agriculture was disparaged by gentry and cultural elites as something suitable only for lower classes. Yet, in the context of evangelical publishing, broadcasting came to be viewed not as a popular practice in opposition to elite preferences, but as a popularizing practice used by elites to aid their purposes—in the case of evangelical publishers, to stem the tide of irreligion and expanding democracy which was regarded by them as mob rule. The shift here is not a reversal, nor is it cumulative. Rather, it derives sense from the earlier agricultural usage, but has been remade to fit the new context, thus emerging from but also severed from the past.

This heuristic also allowed me to recognize transitions, exchanges and developments between the agricultural and media usages. These could be mapped along some key dimensions, which emerged from my theoretical perspective as well as from the documentary evidence.

The first dimension concerns the concreteness of the practice, with a second, related dimension concerning the presence and place of human agency in this process. On the one hand, what could be called a metaphorical usage expresses the natural widespread distribution of non-material qualities or ideas. To the degree that the metaphorical usage removes human agency in any form except as implied recipients of this natural process, I took this as evidence of the mystification of broadcasting. For example, in the metaphorical, mystified usage, writers express “the growth of those [desirable] affections whose seeds are sown at broadcast in the natural relations of life”

(Sedgwick 1843). No mention exists of human intention, decision or action. Similarly, educator Horace Mann expresses in a speech that “the way is open here in Kentucky for sowing the seeds of freedom broadcast” (“Speech of the Hon. Horace Mann” 1848).

On the other hand, what could be called a practical usage names the intentional human action of distributing physical media products, such as newspapers, books, pamphlets, and broadsides. Unlike the metaphorical, the practical usage places people as the instigators and concrete means for a wholly human process. For example, writers advocate that “such notices [for abolition] should be thrown into the tract form, and sown broadcast over the land” (*National Era* 1847, 15 April, 4). And it is asserted that the facts of “the brutal character of American slavery” are available for all to read because they “are sent broadcast over the British empire, and are published in nine-tenths of all the British newspapers” (“Monthly Illustrations of Slavery” 1848).

Yet, by remaining close to the documentary evidence, I resisted the temptation to rigidly equate the metaphorical with the mystified. The resulting view of this process I gained lends credence to Williams’s contention that cultural analysis of this kind reveals a complex and contradictory mixing and melding of different meanings, not simply a linear replacement of one by another. Some items I had found refer to the active distribution of media products (practical), but as a natural process unaided by human intention or intervention, thus mystifying their production, use, and circulation. An example is one statement that notes how documents (practical) seemingly under their own power (metaphorical) “would go broadcast all over the free States in this country [ ... ]” (“Counter Statement” 1840). Others refer to an intentional human activity (practical), but of the spreading of ideas which lacked any material embodiment (metaphorical). People are claimed to “go sowing broadcast the seeds of rich blessing,” “sow also broadcast and with a liberal hand, the seed of truth,” or have “sown broadcast the seeds of bitterness [ ... ]” (“Review of The Young Lady’s Friend” 1838; “Spring!” 1848; “Michigan Politics” 1849).

### *Media broadcasting solidifies*

As I continued to interpret chronologically, the fuzzy distinction between the practical and metaphorical forms of broadcasting slowly began to clarify. With increasing frequency, by the 1850s and 1860s explicit discussions take place regarding the role of media and broadcasting in agitating for an end to slavery. Such items express increasingly clearly a one-way if not vanguard relationship between the activist group and the target of its efforts. Indeed, the clarity of such mentions accelerates throughout the 1850s to the point where broadcasting names and embodies the political work done by media employed by vanguard groups. Whether a metaphorical or a practical process, broadcasting in the service of the social movement comes to mean, for the instigators, an intentional, concrete practice and, for the audiences, reception and acceptance. In one item, “anti-slavery books, tracts, and newspapers should be scattered broadcast over the land; [and] the question of abolition should be brought home to every man’s hearth-stone” (“Resolutions” 1851). Claims of effect become increasingly certain. As one writer puts it, no one can “doubt for a moment the utility of scattering broadcast over our land, such documents and speeches as will have a tendency to enlighten the public mind” (“A Circular to the Friends of the Republican Movement” 1856). An organization suggested that “our publications should be sown broadcast in

all our advances, and keep pace with our advancing missionary operations, as a sure means of permanent success" ("Have We Made Any Advancement" 1864).

By 1859, this sense broadens further. The key tactic of mass social activism had come to be defined not as a specific, practical technology (newspaper, tract, pamphlet) but generally as broadcast distribution, whether applied to newspapers, conversations or debates. The irrelevance of the specific technology in favor of a recognition of the power of broadcasting was increasingly axiomatic. "Scatter information broadcast among the people," urged a writer in 1859 with inflections of the agricultural, "the harvest will be sure" ("Speech of Gov. Chase at Cincinnati" 1859). By the early 1860s, this complex of relationships and practices—a vanguard relation between advocacy groups and their targets, and broadcasting as a reified spread of sentiment as well as a practical tactic of political agitation—were explicitly and uniformly articulated.

### *The usage generalizes*

As I continued to read, I saw how the class specificity of broadcasting and its effectiveness with non-elites also articulated with other class-striated debates in which broadcasting figured prominently. While abolitionists felt increasingly confident concerning its power, broadcasting was not seen by everyone as beneficial or admirable. Those who opposed abolition felt the press "contributed to bring about a ferocious discontent, which needed only the insidious and inflammatory articles spread broadcast over the land by designing men to fan into an insurrection" (Blumenthal 1851).

Opposition to broadcasting also was present in concerns about the moral effects of mass, commercialized popular culture, which had emerged in the form of national markets for popular books. On the one hand, religious elites approved of broadcasting as framed within and linked to the Protestant tradition of direct access to God and the popularization of religious practice, with medieval society dismissed as one in which "the Word of Truth was read only by the learned men and cloistered monks, and its benign influence was not shed broadcast over the earth as now [...]" ("Clerical Humanity" 1851).

Yet, while it was fine to broadcast Protestant Christianity, broadcasting other kinds of materials and sentiments was less so. For example, growing concerns appear about the distribution of French popular novels that are "sown broadcast through our land [...] corrupt[ing] the imaginations of our young men and maidens, and worse [...] wast[ing] the time of more mature readers" ("Editors' Table" 1863). Similar claims appear regarding how "the mails are extensively prostituted to immoral and vicious purposes, and that through this channel obscene books, circulars, &c. are sown broadcast throughout the country" ("Domestic Items" 1863).

Additional research to find primary and secondary sources underscored the value of visual evidence for historical interpretation (Brennen and Hardt 1999). Searching by using the terms "sower" and "broadcast" located important visuals. For example, in a volume of poetry by Bryant (1871), a woodcut neatly depicted the agricultural usage of broadcast seen as a human activity. In it, a bearded male peasant in realist silhouette and placed in a much larger field flings seeds from a bucket hung by its handle in the crook of his arm (Bryant 1871, 23). However, other images in it graphically confirmed broadcasting as a glorification of non-human power and control. The first woodcut image is in high contrast to another woodcut also reproduced in Bryant, which depicts

the millennial, messianic power of broadcasting typical of 19th-Century Christian fundamentalist conceptions. In it, a Medusa-headed androgynous human figure, bent at the waist and hips thrust to one side, squeezes fists full of seeds into whirlwind circulation. Encompassing this figure is a spherical force-field seemingly produced by broadcasting floating a few feet above the ground, with the external world pummeled by a furious storm, complete with lightning bolts in the top left and right corners of the image (Bryant 1871, 22).

Diagrams and photography were also revealing. By the end of the century, broadcasting in a rational, commonsense manner as the one-way delivery of messages is depicted in a pair of schematic drawings reproduced in a popular-science journal and a book published 25 years after it ("The Radiophone" 1899; Hawkes 1923, 170). By the 1920s, efforts to photograph broadcasting commonly centered and filled the frame with a room-sized machine replete with multiple sets of dials and switches, with its human operator tucked to one side, clearly secondary in importance (Dowsett 1924, 54).

### *Extension into the 20th Century*

While the study in its final form continued by analyzing additional primary research that enlarged an understanding of the extent to which the media definition of broadcasting had solidified by the 1920s, it confirmed and extended rather than refuted the insights derived from the research into the 19th-Century transformation of broadcasting. One interesting development was a moderation if not counterbalance in claims about the power of broadcasting subsuming the responses of an audience to the wishes of a vanguard. By the mid-19th Century, a sense appears of an active, shaping response by the audience consistent with orthodox classical-liberal conceptions of the marketplace, such as the author of a broadside who had "sent so many thousands of 'lower law sermons' broadcast throughout the free States." This author must, argued the commentator, "meet the penalties of the popular will" should his argument not be received approvingly ("Congressional Proceedings" 1852).

What here is a political statement about the ideal workings of a democracy as the sovereignty of the people also fits an implicit, parallel claim about the sovereignty of consumers as conceived in classical economic thought, thus taking us to current times. And it is the place of the public that is most provocative in the modern mass-media sense of broadcasting.

By the early 20th Century, broadcasting had become an industrialized process that not only partitioned makers from audiences but that institutionalized a form of sociality paradoxically collective in its pervasiveness but individualized in its experience. Recalling my interest in broadcasting as a complex of social relationships, a social conception of "broadcasting" even more fundamental than a wireless means of disseminating commercial entertainment is lack of reply. It is here that I recognized the full relevance of the media-centric usage of broadcasting that I had found early on in the OED, a facsimile copy of which I subsequently secured. This 1921 item notes that a sea-board wireless station "is used partly for broadcasting Press and other messages to ships, that is, sending out messages without receiving replies" (Crawley 1921, 92). By this time, reply and dialogue in broadcasting are neither expected nor technically possible. Whereas an agricultural usage deems a non-response to be a bad crop (the seeds

didn't germinate and grow), this example suggests that, whether for better or worse, response is wholly and simply irrelevant.

### **Unearthing broadcasting**

What I was able to produce through this analysis—enabled crucially by full-text searches of digital archives of historical periodicals—is an account of the unearthing of broadcasting. It used a theoretically driven inquiry to document how the trajectory of the transmogrification of broadcasting from a human practice of agriculture to the electromagnetic basis of modern media systems is a gradual, sedimented, radically historical process of mystification and naturalization—the transformation of an intrinsically human activity into one of nature and thus by definition fixed, commonsense and thus outside human control and beyond question, critique, or change.

Historians all too infrequently get an opportunity to reflect in detail upon their work. And this is unfortunate. Doing so reminds historians of the theoretical constitution of explanations and of the significance of empirical material, and that these are in mutual relation. Documents do not stand on their own as self-evident proof of an interpretation, just as an explanation does not stand on its own outside of empirical work or theoretical perspective. It is through understanding reflexively the mutual relationships between the empirical, the explanatory and the theoretical where a fully critical–historical method can be fashioned and practiced—and the problems avoided of elevating one of these facets as the ultimate check on accuracy and validity.

In addition to its general epistemological value, critical commentary on media history provides readers with a more granular view. It enhances an understanding of the rationale for choices made, thus opening up one's own research process to additional and detailed reflection and critique. Critical commentary on method also underscores the difficulty of developing innovative approaches, the importance of hunches that sometimes help develop those innovations, as well as the insights that can be gained.

Ultimately, the value of this study is not simply in finding a different way to document claims already made in the voluminous secondary literature on broadcasting. Rather, its contribution is the demonstration of a theoretically driven, evidence-based interpretation aided by the use of digital archives of historical periodicals to unearth how and why this came about.

### **DISCLOSURE STATEMENT**

No potential conflict of interest was reported by the author.

### **REFERENCES**

- "A Circular to the Friends of the Republican Movement." 1856. *The National Era*, 17 January, 12.
- Aitken, Hugh G. J. 2014. *The Continuous Wave: Technology and American Radio, 1900–1932*. Princeton: Princeton University Press.
- Anstruther, John. 1796. *Remarks on the Drill Husbandry*. London: T. Egerton.
- Barnouw, Erik. 1966–70. *A History of Broadcasting in the United States*. 3 vols. New York: Oxford University Press.

- Blumenthal, Charles E. 1851. "Develour; A Sequel to 'The Niebelungen'." *Godey's Lady's Book*, January, 51–54.
- Brecht, Bertolt. 1983. "Radio as a Means of Communication: A Talk of the Function of Radio." In *Communication and Class Struggle, Vol. 2: Liberation, Socialism*, edited by Armand Mattelart and Seth Siegelaub, 169–71. New York: International General.
- Brennen, Bonnie, and Hanno Hardt. 1999. *Picturing the Past: Media, History, and Photography*. Urbana: University of Illinois Press.
- Briggs, Asa. 1985. *The BBC: The First Fifty Years*. Oxford: Oxford University Press.
- Brown, Candy Gunther. 2004. *The Word in the World; Evangelical Writing, Publishing, and Reading in America, 1789–1880*. Chapel Hill: University of North Carolina Press.
- Bryant, William Cullen. 1871. *The Song of the Sower*. New York: D. Appleton.
- "Clerical Humanity." 1851. *The National Era*, 10 July, 110.
- Collins, Edward John T. 1989. "The 'Machinery Question' in English Agriculture in the Nineteenth Century." In *Research in Economic History, Suppl. 5, Pt. A: Agrarian Organization in the Century of Industrialization: Europe, Russia and North America*, edited by George Grantham and Carol Leonard, 203–217. Greenwich and London: JAI Press.
- "Congressional Proceedings." 1852. *The National Era*, 8 January, 1.
- "Counter Statement." 1840. *The Colored American*, 16 May, 2.
- Covert, Catherine L. 1984. "We May Hear Too Much': American Sensibility and the Response to Radio, 1919–1924." In *Mass Media Between the Wars: Perceptions of Cultural Tension, 1918–1941*, edited by Catherine L. Covert and John D. Stevens, 199–220. Syracuse: Syracuse University Press.
- Crawley, C. G. 1921. "Maritime Wireless." *Discovery: A Monthly Popular Journal of Knowledge*, April, 91–94.
- Czitrom, Daniel. 1982. *Media and the American Mind; From Morse to McLuhan*. Chapel Hill: University of North Carolina Press.
- Davison, Lee, Tim Hitchcock, Tim Keirn, and Robert B. Shoemaker, eds. 1992. *Stilling the Grumbling Hive: The Response to Social and Economic Problems in England, 1689–1750*. New York: St. Martin's Press.
- "Domestic Items." 1863. *The Christian Recorder*, 30 May, 91.
- Dowsett, Harry Melville. 1924. *Wireless Telephony and Broadcasting*. Vol. 1. London: Gresham Publishing.
- "Editors' Table; Novels in French and English." 1863. *Godey's Lady's Book*, March, 304.
- Fussell, George Edwin. 1947. *The Old English Farming Books: From Fitzberbert to Tull, 1523 to 1730*. London: Crosby Lockwood & Son.
- Gitelman, Lisa, and Geoffrey B. Pingree, eds. 2003. *New Media, 1740–1915*. Cambridge and London: MIT Press.
- Hamilton, James F. 1991. "Editorial and Introduction." *Journal of Communication Inquiry* 15 (2): 5–11.
- Hamilton, James F. 2007. "Unearthing Broadcasting in the Anglophone World." In *Residual Media*, edited by Charles Acland, 283–300. Minneapolis: University of Minnesota Press.
- Hamilton, James F. 2008. *Democratic Communications: Formations, Projects, Possibilities*. Lanham: Lexington Books.
- Hardt, Hanno, and Bonnie Brennen. 1993. "Introduction: Communication and the Question of History." *Communication Theory* 3 (2): 130–136.
- "Have We Made Any Advancement?" 1864. *The Christian Recorder*, 4 June, 90.



- Hawkes, Ellison. 1923. *The Romance and Reality of Radio*. London: T. C. & E. C.
- Hiken, Louis. 1999. "Forward." In Lawrence Soley, *Free Radio: Electronic Civil Disobedience*, ix–xii. Boulder: Westview Press.
- Hilliard, Robert L., and Michael Keith. 2001. *The Broadcast Century and Beyond: A Biography of American Broadcasting*. 3rd ed. Boston: Focal Press.
- Hilmes, Michele. 1990. *Hollywood and Broadcasting: From Radio to Cable*. Urbana and Chicago: University of Illinois Press.
- Mander, Mary, 1984. "The Public Debate About Broadcasting in the Twenties: An Interpretive History." *Journal of Broadcasting* 28 (2): 167–185.
- Marvin, Carolyn. 1987. *When Old Technologies were New: Thinking About Communications in the Late Nineteenth Century*. New York: Oxford University Press.
- Marx, Karl, 1967. "The Fetishism of Commodities and the Secret Thereof." In *Capital, Vol. 1*, 71–83. New York: International Publishers.
- McCauley, Michael P. 2002. "The Contested Meaning of Public Service in American Television." *The Communication Review* 5: 207–237.
- "Michigan Politics." 1849. *The National Era*, 25 October, 169.
- Mingay, G. E. 1994. *Land and Society in England 1750–1980*. London and New York: Longman.
- Mingay, G. E. 2002. *A Social History of the English Countryside*, reprint ed. London and New York: Routledge.
- "Monthly Illustrations of Slavery." 1848. *The North Star*, 25 August, 2.
- Myers, Cayce, and James F. Hamilton. 2014. "Social Media as Primary Source: The Narrativization of 21st-Century Social Movements." *Media History* 20 (4): 431–444 doi:10.1080/13688804.2014.950639.
- Nord, David Paul. 1984. "The Evangelical Origins of Mass Media in America, 1815–1835." *Journalism Monographs* 88.
- Nord, David Paul. 2004. *Faith in Reading; Religious Publishing and the Birth of Mass Media in America*. New York: Oxford University Press.
- OED Online. 2017. "broadcast, v." Oxford University Press. <http://www.oed.com.proxy-remote.galib.uga.edu/view/Entry/23508> (accessed January 11, 2018).
- Peters, John Durham. 1999. *Speaking into the Air; A History of the Idea of Communication*. Chicago: University of Chicago Press.
- "Resolutions Adopted by the late Anti-Slavery Convention in Cincinnati." 1851. *The National Era*, 15 May, 79.
- "Review of The Young Lady's Friend." 1838. *The Lady's Book*, April, 167–169.
- Scannell, Paddy. 1991. "Introduction: The Relevance of Talk." In *Broadcast Talk*, edited by Paddy Scannell, 1–13. Newbury Park: Sage.
- Sedgwick, C. M. 1843. "Scenes from Life in Town." *The Lady's Book*, April, 159–163.
- Slack, Jennifer Daryl, and J. MacGregor Wise. 2002. "Cultural Studies and Technology." In *Handbook of New Media; Social Shaping and Consequences of ICTs*, edited by Leah Leivrouw and Sonia Livingston, 485–501. London: Sage.
- "Speech of Gov. Chase at Cincinnati." 1859. *The National Era*, 10 November, 180.
- "Speech of the Hon. Horace Mann." 1848. *The National Era*, 21 September, 150.
- Spinelli, Martin. 2000. "Democratic Rhetoric and Emergent Media; the Marketing of Participatory Community on Radio and the Internet." *International Journal of Cultural Studies* 3 (2): 268–278.
- "Spring!" 1848. *The North Star*, 14 April, 2.

- "The Slave Trade." 1847. *The National Era*, 15 April, 4.
- "The Radiophone at the Electrical Exhibition." 1899. *Scientific American*, 27 May, 347.
- Thorburn, David, and Henry Jenkins, eds. 2003. *Rethinking Media Change; the Aesthetics of Transition*. Cambridge and London: MIT Press.
- Williams, Raymond. 1976. *Keywords: A Vocabulary of Culture and Society*. New York: Oxford University Press.
- Williams, Raymond. 1977. *Marxism and Literature*. Oxford: Oxford University Press.
- Williams, Raymond. 1980. "Base and Superstructure in Marxist Cultural Theory." In *Problems in Materialism and Culture*, edited by Raymond Williams, 31–49. London: Verso.
- Williams, Raymond. 1989. "Communications, Technologies and Social Institutions." In *What I Came to Say*, 172–192. London: Hutchinson Radius.



# EXPLORING MACHINE LEARNING TO STUDY THE LONG-TERM TRANSFORMATION OF NEWS

Digital newspaper archives, journalism  
history, and algorithmic transparency

Marcel Broersma  and Frank Harbers 

*The labour-intensive nature of manual content analysis and the problematic accessibility of source material make quantitative analyses of news content still scarce in journalism history. However, the digitization of newspaper archives now allows for innovative digital methods for systematic longitudinal research beyond the scope of incidental case studies. We argue that supervised machine learning offers promising approaches to analyse abundant source material, ground analyses in big data, and map the structural transformation of journalistic discourse longitudinally. By automatically analysing form and style conventions, that reflect underlying professional norms and practices, the structure of news coverage can be studied more closely. However, automatically classifying latent and period-specific coding categories is highly complex. The structure of digital newspaper archives (e.g. segmentation, OCR) complicates this even more, while machine learning algorithms are often a black box. This paper shows how making classification processes transparent enables journalism scholars to employ these computational methods in a reliable and valid way. We illustrate this by focusing on the issues we encountered with automatically classifying news genres, an illuminating but particularly complex coding category. Ultimately, such an approach could foster a revision of journalism history, particularly the often hypothesized but understudied shift from opinion-based to fact-centred reporting.*

## Introduction: From Scarcity to Abundance

Access to old news has been improved tremendously in the past decades. National libraries in for example France, Australia and the Netherlands have digitized their historical newspaper collections on a large scale while many local archives have digitized individual titles that cater to the interests of regional historians. In contrast to this “public model” which provides free access to everyone, there is a “commercial model” that has been applied in countries such as the US and the UK. Here publishers

have sold their rights to commercial companies such as Cengage, which have digitized papers and created databases to sell subscriptions to universities, public libraries and archives (cf. Nicholson 2013; Mussell 2008), hampering the access to research material. Nevertheless, cost considerations and accessibility issues aside, research can now be done from behind the desk and is thus less time-consuming.

Digitization has clearly opened up new venues for journalism research because large text corpora are now full-text searchable. However, the quality and availability of historical newspaper archives diverges considerably, and with that their value for research. First, access to sources differs between countries and between archives. While public archives might under special conditions be inclined to provide researchers access to the complete data set of text files (and metadata), commercial companies are hesitant because it jeopardizes their business model. Second, because of copyright issues periodicals are often not available after 1945. This puts severe restraints on comparative and longitudinal research. Third, the enthusiastic uptake of large-scale digitization projects and the advancement of technology have resulted in public, semi-public and commercial silos using different digitization standards and procedures. This not only prohibits data integration on a higher level, but also results in different quality of text files due to issues with OCR and segmentation. The number of errors in text files generated in the digitization process also differs tremendously between historical periods and publications (cf. Broersma 2011a, 2011b; Wijfjes 2017).

Even more importantly, these archives generally only provide keyword search possibilities, based on simple and more complex queries applying for example (Boolean) operators and wildcards, as the main gateway to their content. Therefore, data can usually only be searched and retrieved through the search interface, resulting in a list of individual hits. While this is a major step forward compared to endlessly scrolling through microfilms or turning pages, it remains unsatisfying. Keyword search only affords straightforward queries, as Deacon (2007, 8) argues: "key word searching is best suited for identifying tangible 'things' (i.e. people, places, events and policies) rather than 'themes' (i.e. more abstract, subtler and multifaceted concepts)." Moreover, metadata, which can be used to limit search results, tend to be added sparsely.

The keyword search tools and the opportunity to download pages in PDF-format most archives offer, are usually sufficient to cater to the demands of the general audience and scholars who consult historical newspapers as a source of specific historical information about a certain event or issue. However, it limits more specialized data publics such as journalism scholars and media historians, who want to study newspapers as a serial source and are interested in the structural transformation of news and journalism. The key question is, therefore, whether and how digitization will actually change research practices in journalism history (cf. Boumans and Trilling 2016; Flaounas et al. 2013). Despite the development of (computer-assisted) social scientific ways of research such as (automatic) quantitative content analysis that offer the opportunity to explore news content beyond ideographic and myopic studies, journalism historians have been reluctant in adopting quantitative and computational methods (Wijfjes 2017; Nicholson 2013; Broersma 2011a, 2011b).

We join in calls for journalism scholars to move beyond keyword search and manual content analysis and take full advantage of the available digitized newspaper material (Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016;

Jacobi, Van Atteveldt, and Welbers 2016; Burscher, Vliegenthart, and de Vreese 2015). Computational methods based on machine learning enable us to root analyses in large data sets instead of necessarily modest samples (Boumans and Trilling 2016; Broersma 2011a, 2011b; Wijfjes 2017). This implies that “we no longer have to choose between data size and data depth” (Manovich 2012, 466). Such approaches not only allow us to ask new questions, but also to come to new conclusions—and challenge the ones not rooted in textual analysis or just based on small subsets of newspaper content. Automatic content analysis, e.g. text statistics, sentiment analysis, topic modelling or frame analysis, facilitates detailed analyses of newspapers as a serial source on an unprecedented scale in a much more cost-efficient way (cf. Boumans and Trilling 2016).

Successfully implementing advanced digital methods could help to systematically study more conceptual questions such as the historical shift of topics and concepts. In addition, we argue for taking a next step by automating content analysis of the formal structure of texts and images. Enabling and facilitating such longitudinal analyses would address an important and peculiar gap in journalism history. Until now, due to methodological and practical issues, research has spent only limited attention to content analysis of the formal characteristics of news (cf. Bingham 2010). The time-intensive nature of this kind of research and the accessibility of analogue newspaper collections put up too high barriers. Though understandable from a practical perspective, this neglect is nevertheless remarkable since the value of journalism for society is first and foremost based on its capacity to provide legitimate representations of social reality, for which form is a crucial category as Broersma (2011b) has argued.

Only recently, scholars, often in interdisciplinary teams of historians, programmers and data scientists, have started to tap into the vast collections of digitized historical news texts. We agree with the growing body of literature on computational methods in journalism research that forms of automated content analysis, specifically machine learning approaches, offer promising venues to analyse big data sets of news content and introduce new questions and approaches to journalism studies (Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016; Jacobi, van Atteveldt, and Welbers 2016; Burscher, Vliegenthart, and de Vreese 2015). It allows for grounding analyses in big data and mapping the structural transformation of journalistic discourse on a large scale.

Still, current approaches mostly focus on recent rather than historical news texts. Digital born data sets are easier to gather, contain less to zero (OCR) errors, and do not have to account for change over time. Furthermore, the emphasis tends to be on the topical content of news texts. The frequency, variety or co-occurrence of words are used as manifest indicators of topics, frames or sentiments in a “stable” synchronic data set (see for instance, Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016; Burscher et al. 2014). While important, this does not provide insights into the *structural transformation* of news and journalism, which sheds light on how journalism “works” beyond day-to-day news stories. While the content of news changes every day, form and style are more stable categories. They indicate professional norms about how journalism needs to be performed and what accounts for a truthful and trustworthy representation of reality. The next step is thus to further develop methods and tools that allow us to analyse historical, diachronic data sets to map the development of news and journalism.

In this article, we argue that a focus on genre, as a marker of professional ideology, enables us to gain important new insights into the development of journalism. Textual characteristics relating to journalistic form conventions and modes of expression, such as genre, are until now rarely discussed or problematized thoroughly in scholarship (cf. Boumans and Trilling 2016). Moreover, we discuss why supervised machine learning is a fruitful and promising approach to automating genre classification and outline how we have operationalized this for our research. We also discuss the issues that complicate the automation of such a complex latent content category. Pivotal in this discussion is the importance of creating a transparent assessment of the performance of machine learning algorithms—too often an opaque black box process. This discussion reveals how in these highly complicated machine-learning tasks transparency is imperative in moving historical research forward.

### **The Neglect of Newspaper Form in Journalism History**

In historical research, most studies that use historical newspaper material do not study or critically reflect on the medium itself and its development. Newspapers are still mainly used to get factual information about historical figures, events and issues—often only to add flavour with telling citations. The medium-specific qualities and its consequences for the source material are often left out of the equation. In this sense not much has changed since historian Brian Maidment argued in 1990 that scientific progress could only be made “if we regard periodicals not like fossil hunters, in search of specimens to fill a cabinet, but like theoretical geologists or theologians, as expositions of processes by which change occurs and is made legible” (quoted in: Vella 2009, 205).

With notable exceptions (Barnhurst and Nerone 2001; Barnhurst 2016; Fink and Schudson 2014), quantitative analyses of news content and textual conventions, inspired by research approaches in the social sciences and aimed at theory building by tracing patterns, remain very scarce in journalism history (see for more details: Broersma 2011a). The history of journalism has largely been studied through archival research into journalism’s institutional development and the analysis of discourses on journalism, such as public statements, debates and autobiographical writing of journalists. These are taken at face value rather than that the strategic nature of such discourses is critically studied (Broersma 2010a).

This has resulted in a distorted picture of the historical development of journalism (Nerone 2010; Harbers 2014; Broersma 2018). Historians have created a transnational grand narrative, which has become known as the “liberal narrative” (Curran 2009). It frames the development of journalism since the nineteenth century as a linear development from a partisan press to an independent and autonomous press. It emphasizes “the establishment of an autonomous profession that, independent from political and economic powers, obeys more or less to the objectivity regime, and the practices and formal conventions resulting from it” (Broersma 2018; cf. Broersma 2007, 2010b; Harbers 2014). However, longitudinal content analysis suggests that this dominant narrative is actually skewed and overemphasizes the innovative nature and pace of journalistic development (Harbers 2014). Such research reveals what Dahlgren (1992, 7) called “the gap between the realities of journalism and its official presentation of self.”

In addition to qualitative textual analysis, quantitative content analysis that traces the development of the textual characteristics of newspaper content can add nuance and complexity to our picture of journalism history. Analysing a representative sample of daily news coverage can elucidate how journalism's modes of expression were employed in everyday practice as well as their historical transformation. It abstracts from the specific content of a news item to broader predefined categories that are traced over time. Specifically, a focus on formal conventions—how a news item is structured and written, and how it is presented to readers—allows us to move beyond day-to-day news events and tap into the underlying structure of news coverage (Broersma 2010b).

These textual conventions pertaining to journalism's form, i.e. the arrangement of layout, genre, and narrative structure and devices, allude to journalism's professional norms and broader cultural discourses (Broersma 2007, 2010b). Such textual conventions can thus reveal "the way[s] the medium imagines itself to be and to act. In its physical arrangement, structure, and format, a newspaper reiterates an ideal for itself" (Barnhurst and Nerone 2001, 3). Studying these formal characteristics historically sheds light on how journalism's modes of expression have gradually transformed and how professional norms and practices, such as the objectivity regime, have emerged and evolved (Broersma 2010a; Benson 2005).

### Genre as Marker of Journalistic Style

For our computational approach to the long-term transformation of journalism, we focus on genre as a marker for professional ideology. Genre is an important characteristic of the form of news content. It structures news discourse and signals to readers what they can expect of an article. Specific genres have been invented throughout journalism history to contain new modes of reporting, reflecting the underlying professional ideology. For example, the report, a prolific genre throughout the 19th century that registered meetings and events chronologically and almost verbatim, was replaced in the 20th century by genres such as the reportage, features and the interview that reflect active reporting and highlight the autonomy of the reporter as interpreter of events (Broersma 2007, 2008, 2010b).

Hartsock (2000) makes a useful distinction here between "topical genres" and "modal genres." This reflects an important difference between Anglo-American and European genre conventions. Within the first cultural context genre refers to a *practice* focused around a certain *topic* or "beat." From this perspective any journalistic texts focused on, for example, sport belongs to the genre "sports journalism." In contrast, modal genres—which we study—refer to a set of *formal conventions*, i.e. particular ways of *structuring* texts, which cut across topics. A news article, for example, would be considered a genre. In this case the inverted pyramid model can be regarded as a typical characteristic of how a news report is structured and thus helps to identify this particular genre. An interview is a genre that centres on a conversation between two people and can be discerned by the way the text is structured around questions and answers. In European journalism, such modal genres are considered the cornerstone of the profession. They are central to the training of aspiring journalists and dealt extensively with in textbooks and journalism programs. Students are trained to know the

differences between genres as these are typically used when assigning stories in newsrooms (Broersma 2008).

We define genre as “language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms” (Bhatia 2004 cited in Handford 2010, 258). As such, the rise and use of particular modal genres indicate the identity of newspapers as they refer to certain styles of journalism. By studying genre conventions, we can, therefore, gain insight into the underlying discursive context: the “communicative goals” of journalism and the professional norms and practices they see as necessary to achieve those goals (Broersma 2010b). The interview and the reportage, for instance, only emerged and became popular from the 1880s onwards and are clearly linked to the shift from reflective, opinion-oriented journalism to fact-based and event-centred reporting (Broersma 2008; Harbers 2014). Examining to what extent these genres are used in newspapers and tracing this over time, therefore, offers important insights into the way journalism was conceived and practiced and how it developed historically (Broersma 2007, 2010b; Harbers 2014).

### **From Manual Content Analysis to a Machine Learning Approach**

In our large-scale research project “Reporting at the Boundaries of the Public Sphere. Form, Style and Strategy of European Journalism, 1885–2005” manual quantitative content analysis offered valuable insights.<sup>1</sup> To study the long-term transformation of news we conducted a longitudinal content analysis of nine European newspapers, which resulted in a database with the metadata, i.e. the topic and genre label, amount and type of sources and images and quotation patterns of 125.000 articles. Yet, this is a highly time consuming and, therefore, expensive endeavour. Moreover, the size of the material that can be annotated is still only a small percentage of the total amount of available material (Harbers 2014).

In line with the potential for journalism studies as outlined by Boumans and Trilling (2016), we see automatic content analysis as highly suitable for longitudinal and comparative research into the historical development of newspapers. However, so far it has mostly focused on text mining and topic modelling (Lee and Myaeng 2002). Studying what kind of topics newspapers reported on in certain periods is certainly fruitful as it can reveal the commercial and ideological strategies of publishers who cater to the demands of different audiences, or show how news values have changed over time (Yang, Torget, and Mihalcea 2011). Nevertheless, it does not offer information on how these topics are presented to the readership or if articles adhere to a fact-based or opinion-oriented style of journalism. The formal characteristics of newspaper texts, such as genre, are still largely left aside. To get a more fine-grained analysis of the underlying journalistic conceptions and modes of expression that newspapers have employed throughout history, it is necessary to move to latent categories of analysis relating to form and style conventions.

Genres, and form conventions in general, are complex latent variables with many and complex sub-categories that are hard to operationalize. Classification of such variables requires much interpretation because they relate to concepts that cannot be observed at the surface of the text, “but can be represented or measured by one or

more [...] indicators" (Hair et al. 2010 cited in Neuendorf 2002, 23). Human coders, therefore, need extensive training to reach acceptable inter coder reliability levels (Neuendorf 2002; Harbers 2014). Doing this automatically proves to be an even more difficult task for computers. As Manovich (2015, 22) explains, forms of automatic content analysis have to deal with a "semantic gap." Machines only recognize manifest "low level information." In the case of newspaper articles, this means semantic and lexical characteristics as well as punctuation marks—though the latter are generally excluded during the preprocessing phase (Günther and Quandt 2016).

Texts are often represented as "bags of words," showing the frequency of each word but disregarding their order. Lexical features are mostly used to disregard less interesting types of words, such as prepositions, articles, and adverbs, and zoom in on (proper) nouns (Jacobi, van Atteveldt, and Welbers 2016). Representing the text through these characteristics is very suitable for determining its topic or frame as it displays the semantic particularities of a text directly related to its meaning (Günther and Quandt 2016; Burscher et al. 2014). However, this approach is too limited for classifying formal characteristics such as genre as it does not include textual characteristics like quotes, metaphors or narrative structure and perspective that allude to the form of texts. These are much harder to operationalize in such a way that a machine can distinguish them.

This makes automating research into formal characteristics of texts such as genre an extremely complicated task. Genres cut across topics and, therefore, cannot be discerned based on semantic characteristics—or at least not solely. Take for instance a reportage. This genre not only provides factual information, but also conveys the atmosphere and the experience of witnessing a certain event or issue, which can range from politics to war, sports and lifestyle. Articles are often structured chronologically, depict a detailed picture of the space and surroundings, convey what the reporter and her sources saw and felt, and use imagery to make the experience of "being there" tangible. In addition, contextual indicators such as the article's length, self-classification and position within the newspaper provide cues.

Such characteristics need much human interpretation and are hard to translate to the low level textual characteristics a computer algorithm can deal with (cf. Manovich 2015; Boumans and Trilling 2016). To complicate things even more, genres are ideal-typical discursive constructs. This means that textual manifestations do not always match the characteristics of these constructs perfectly. Deciding when articles that only partially comply with the textual characteristics of a certain genre can still be considered representative of that genre is challenging. In addition, articles might share characteristics with other genres and genres are dynamic constructions that change or fade away over time while new ones emerge. This makes it hard to rigidly define genres as is necessary for quantitative and deductive forms of (automatic) content analysis (Harbers 2014). To deal with the historical variety of texts within different genres two approaches can be taken. Ideally, an algorithm would be trained on the basis of a training set that covers the entire historical period and recognizes the different genres across history. However, at the moment, this might still result in a low accuracy level and an alternative approach is to train different algorithms based on different training sets for different historical periods.

As such, classifying genres of historical newspaper articles with appropriate levels of validity is clearly challenging. However, we argue that (supervised) machine learning offers a promising approach compared to more traditional ways of automatic content analysis. The latter are closely related to the rule-based approach of manual content analysis in which texts are classified according to predefined categories (Lewis, Zemith, and Hermida 2013; Apté, Damerau, and Weiss 1994). Dictionary- and rule-based automatic content analysis operates in a similar way: the machine assigns a label to a text based on the presence of certain textual characteristics, based on lists of specific words or combinations of words relating to a particular topic, theme or concept. This is a rigid and static way of assigning texts to categories (Günther and Quandt 2016; Zamith and Lewis 2015). Moreover, it is very hard to create and validate exhaustive lists as “most people do not know the complete set of words that indicate a particular content category and/or all ways such words can be used” (Burscher 2016, 21).

While a dictionary approach works for certain tasks (such as determining named entities like sources, or the sentiment of an article), it is problematic for automatic classification of formal characteristics of texts. Here, creating a list of particular words is unlikely to work since these characteristics are independent of the content of an article. Supervised Machine Learning (SML) takes a much more open and dynamic approach because the decision process is not predefined. It uses a manually annotated data set based on predetermined coding categories as initial input. This annotated data set is used to develop and train a self-learning algorithm that creates its own discriminatory model to predict which category is the most likely match.<sup>2</sup> A subset of the training data is used to formally evaluate the performance of the algorithm. As such, it also validates the model it uses to predict the genre of a text. After the algorithm is trained and has reached a satisfactory accuracy level, it can be used to classify new texts. In theory, this makes the need for sampling redundant as the entire corpus could be analyzed (cf. Günther and Quandt 2016; Boumans and Trilling 2016; Grimmer and Stewart 2013; Burscher 2016).

### **Applying Machine Learning to Historical Newspaper Archives**

For this kind of research expertise from various disciplines is imperative. In two research projects, we as domain specialists in journalism history, therefore, work closely with collection specialists from different archival institutions, and data and computer scientists.<sup>3</sup> A main challenge is to translate research questions about journalism history to computer science approaches in machine learning. Not only is the past a foreign country, but those who try to map and analyse it also speak different languages. We build on our experiences with these ongoing projects to discuss the opportunities, pitfalls and problems by applying supervised machine learning to automatic genre classification. In addition, we argue for the importance of algorithmic transparency; scholars without computer science expertise should be able to assess the performance of algorithms beyond mere accuracy percentages.

Underlying our research is a manually annotated data set that has been developed in our previous research project into the transformation of European journalism between 1885 and 2005. Although it also contains metadata about French and British newspapers, we only used the genre classifications of a large sample ( $N=33.000$ ) of



Dutch historical newspaper articles, as we only had access to the Dutch corresponding digitized newspaper content.<sup>4</sup> This subset is used to train and evaluate different off-the-shelf machine learning algorithms to see which performs best. Ultimately, this allows us to make an informed choice about which algorithm is most suitable for doing a specific machine learning task.

Other than a dictionary-based or a traditional rule-based approach, machine learning algorithms can deal with and combine many different features of texts and independently decide which ones are relevant to base classifications on. In our project we trained several supervised machine learning methods, such as Support Vector Machines, Naïve Bayes, and Random Forests, to evaluate and compare their performance. While in our manual content analysis human coders reached an accuracy score of 85 per cent for coding genre with corresponding Krippendorff's alpha of 0.83 (Harbers 2014), the automatic classifiers assign the right genre in between 41 and 70 per cent of the cases.<sup>5</sup> Although the best performing algorithms provide promising scores given the very complex task of genre classification, it leaves the reliability of these tools still under par. They need to perform at least above the lowest accepted accuracy score for human intercoder reliability ( $K\alpha$  0.67) to draw robust conclusions (Riffe, Lacy, and Fico 2005). That being said, new experiments and adding more training data are likely to improve the reliability of automatic genre classification consistently (see Bilgin et al. 2018, for a more elaborate discussion of our approach, experiments and evaluation of different machine learning algorithms).

However, a major issue with only assessing these overall accuracy scores is that they are the result of a black box process that obscures the built-in choices and biases that result from the training of an algorithm. Without insight into how an algorithm operates and assigns a genre label, it is impossible to evaluate its validity. This is important because the algorithm that performs best in terms of accuracy does not necessarily label texts in a valid way. Therefore, it is crucial to compare different algorithms, elucidating their built-in choices and biases in a way that makes the strengths and weaknesses of their performance transparent. In our collaborative project, we are, therefore, developing a virtual workspace—a dashboard—in which researchers can do experiments with different algorithms on the same training set. This dashboard enables scholars to explore and compare the performance of algorithms and test the influence of different discriminatory features. Through different visualization techniques available on the dashboard, this approach elucidates how these algorithms function in a way that is comprehensible for end users with a humanities and social science background and allows for an informed evaluation of the best functioning algorithm for a specific goal.

One particular issue we encountered is the skewed distribution of genre categories in newspapers—and, therefore, also in the training set. Some genre categories, particularly news reports, are disproportionately present in newspaper content, whereas genres such as the interview or the reportage appear less frequently. While it is advised to train algorithms based on a representative training set, this approach can lead to “overfitting” or “overtraining.” The algorithm is then likely to show a preference for genres that are overrepresented in the training data. As such, the average performance might look fine, but the algorithm is likely to underperform in identifying underrepresented genres (cf. Zheng, Wu, and Srihari 2004). In selecting a particular machine

learning algorithm, it is, therefore, adamant to be able to evaluate the performance for each genre category. Our comparison showed that the algorithm with the highest overall accuracy score differed considerably in its performance per genre category, e.g. almost perfect on classifying news reports, but very poorly on classifying a reportage. Based on this particular bias, it is a less likely choice for the task it needs to do.

Another issue relates to the specific features of the discriminatory model algorithms use to classify the genre of a text. As we are interested in modal genres and, therefore, the particular way information about a certain topic is communicated, we consider it crucial that the algorithms base their classifications on "modal cues" rather than "topical cues." For this we used "feature importance ranking plots" to see which textual characteristics were deemed relevant in classifying a certain genre. This showed that although some of the algorithms did indeed use modal cues to assign certain genres, such as the reportage, the algorithms displaying the highest level of overall accuracy also show the highest tendency to base their classifications on topical cues. This conflicts with the conceptualization of genre as a category that cuts across topical boundaries.

## Conclusion

Grounded in our experiences with both manual content analysis and supervised machine learning approaches to automatic content analysis, we have argued that the latter offers promising venues for exploring journalism history. We have discussed how such computational methods promise to enable the exploration of large data sets without compromising the depth of the analysis, thus allowing for fine-grained comparative and longitudinal research into the history of news and journalism. We propose to move beyond the relatively easy automation of manifest coding categories (e.g. counting numbers of articles containing certain keywords, or word frequencies within articles) and automatic content analysis based on topic modelling (including frame analysis), because latent content categories pertaining to form, such as genre, can provide us with a more fundamental and detailed insight into the structural transformation of journalism. In our focus on a very complex machine learning task, classifying the latent variable of genre, we compare and evaluate the overall performance of different classifiers, while at the same time rendering the bias and operations of algorithms transparent. Such an endeavor reveals not only the opportunities of our computational approach, but also the persistent problems that need to be solved to fully exploit the research possibilities that digital newspaper archives offer.

At the moment, important progress is made in creating more sophisticated ways to automate the time-consuming method of content analysis, specifically using (supervised) machine learning. To reach valid results, this requires a manually constructed training set with clearly defined categories that can be translated to machine-readable textual features, because, as Simon (2001, 87) argued, "the computer is simply unable to understand human language in all its richness, complexity, and subtlety as can a human coder." However, apart from practical issues concerning the structure and quality of the digital archive, before the potential of a supervised machine learning approach can become reality, important questions about the most suitable procedures to train an algorithm, which algorithm performs best, what the biases are of different

possible algorithms, need to be addressed. Only through enhancing transparency about the training process and performance of such algorithms, we can move toward a trustworthy and reliable approach of analysing the formal characteristics of historical newspaper material.

We consider these approaches to automate forms of content analysis as important additions to the research toolbox of media historians. These will be particularly helpful to test hypotheses that have been formulated based on qualitative research on a larger scale. Similarly, they can be used to map broader developments and trace patterns in historical development that can be contextualized, fleshed out and elaborated by close reading and archival research. In our manual content analysis we could demonstrate how the shift in the course of the twentieth century from opinion-based to fact-centred reporting was far more gradual and messy than is often argued in historical scholarship. Our automated analysis can confirm and refine this conclusion based on a far bigger data set that includes more newspapers while also being distributed more equally over time. Combining computational methods with more contextual qualitative approaches is crucial here because extensive knowledge of the relevant media historical context is pivotal to provide sound and meaningful interpretations of the data generated by algorithms. As Lewis, Zamith, and Hermida (2013, 48) argue: "In the allure of computational methods, researchers must not lose sight of the unique role of humans in the content analysis process. This is particularly true of their ability to bring contextual sensitivity to the content, its manifest and latent characteristics, and its place within the larger media ecology."

In the end, it is important to recognize that automating genre classification, and forms of content analysis in general, should not be regarded as the final solution to the issues with conducting this type of research. As Boyd and Crawford (2012, 671) caution, "context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model." Despite its potential, this approach—and similar approaches—will not replace traditional forms of media historical scholarship, but much rather complement them.

## NOTES

1. This NWO-VIDI project (PI: Broersma) was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 276-45-002. For more details, see the PhD thesis of Harbers who conducted one of the subprojects (Harbers 2014).
2. As we are interested in assigning predefined genre labels that mirror the genres that were current in journalism history we leave forms of unsupervised machine learning aside. For an overview of the opportunities and benefits of unsupervised machine learning for journalism studies, see for instance: Boumans and Trilling 2016; Jacobi, Van Atteveldt, and Welbers, 2016.
3. The first project, "Discerning Journalistic Styles," was conducted by Harbers during a digital humanities fellowship at the Dutch National Library (KB) in 2016 and was a first pilot research that, in collaboration with a data scientist, explored ways to automate genre classification of historical newspaper articles: <http://lab>.

kb.nl/tool/genre-classifier#introduction The second project, “News Genres: Advancing Media History by Transparent Automatic Genre Classification (NEWSGAC; PI: Broersma, Co-applicant: Harbers)” is a bigger follow-up project, funded by CLARIAH/NWO and the Netherlands e-Science Center under project number ADAH.2016.020. It is a collaboration between the Centre for Media and Journalism Studies of the University of Groningen, the national research institute for mathematics and computer science in the Netherlands CWI, the National Library of the Netherlands and the Netherlands Institute for Sound and Vision: <https://www.esciencecenter.nl/project/newsgac>.

4. The National Library of the Netherlands graciously granted us access to their dataset of digitized newspapers that is the result of a large-scale newspaper digitization program running since 2006.
5. The default accuracy by predicting the majority class or genre (news report) is 46%.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

## FUNDING

This work was supported by CLARIAH/NWO and the Netherlands e-Science Center [Grant number: ADAH.2016.020].

## ORCID

Marcel Broersma <http://orcid.org/0000-0002-7342-3472>

Frank Harbers <http://orcid.org/0000-0003-1578-7582>

## REFERENCES

- Allen, Robert B., Ilya Waldstein, and Weizhong Zhu. 2008. “Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres.” In *Digital Libraries: Universal and Ubiquitous Access to Information*, edited by George Buchanan, Masood Masoodian, and Sally Jo Cunningham, 379–386. New York: Springer.
- Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. 1994. “Automated Learning of Decision Rules for Text Categorization.” *ACM Transactions on Information Systems* 12 (3): 233–251.
- Barnhurst, Kevin. 2016. *Mr. Pulitzer and the Spider: Modern News from Realism to the Digital*. Urbana: University of Illinois.
- Barnhurst, Kevin, and John Nerone. 2001. *The Form of News. A History*. New York: Guildford Press.
- Benson, Rodney. 2005. “Mapping Field Variation: Journalism in France and the United States.” In *Bourdieu and the Journalistic Field*, edited by Rodney Benson and Eric Neveu, 85–112. Cambridge: Polity Press.

- Bilgin, Aysenur, et al. 2018. "Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History." Paper Submitted to the 14<sup>th</sup> IEEE eScience Conference, Amsterdam. [Under review]
- Bingham, Adrian. 2010. "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." *Twentieth Century British History* 21 (2): 225–231.
- Boumans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit." *Digital Journalism* 4 (1): 8–23.
- Boyd, Danah, and Kate Crawford. 2012. "Critical questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–679.
- Broersma, Marcel. 2007. "Form, Style and Journalistic Strategies." In *Form and Style in Journalism. European Newspapers and the Representation of News. 1880-2005*, edited by Marcel Broersma, ix–xxix. Leuven: Peeters.
- Broersma, Marcel. 2008. "The Discursive Strategy of a Subversive Genre." In *Vision in Text and Image: The Cultural Turn in the Study of Arts*, edited by Mary Kemperink and Herman, 143–158. Leuven: Peeters.
- Broersma, Marcel. 2010a. "De Transformatie van het Journalistieke Veld: Discursieve Strategieën en Journalistiek Vormen." *Tijdschrift voor Communicatiewetenschap* 38 (3): 267–275.
- Broersma, Marcel. 2010b. "Journalism as a Performative Discourse. The Importance of Form and Style in Journalism." In *Journalism and Meaning-Making: Reading the Newspaper*, edited by Verica Rupar, 15–35. Cresskill: Hampton Press.
- Broersma, Marcel. 2011a. "Nooit Meer Bladeren. Digitale Krantenarchieven als Bron." *Tijdschrift voor Mediageschiedenis* 14 (2): 29–55.
- Broersma, Marcel. 2011b. "From Press History to the History of Journalism. National and transnational features of Dutch scholarship." *Medien & Zeit* 26 (3): 17–28.
- Broersma, Marcel. 2018. "Americanization, or: The Rhetoric of Modernity. How European Journalism Adapted US Norms, Practices and Conventions." In *The Handbook of European Communication History*, edited by Klaus Arnold, Paschal Preston, and Susanne Kinnebrock. Chichester and Malden: Wiley.
- Burscher, Björn, Daan Odijk, Rens Vliegthart, Maarten de Rijke, and Claes H. de Vreese. 2014. "Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis." *Communication Methods and Measures* 8 (3): 190–206.
- Burscher, Björn. 2016. "Machine Learning-Based Content Analysis: Automating the Analysis of Frames and Agendas in Political Communication Research." PhD diss., University of Amsterdam.
- Burscher, Björn, Rens Vliegthart, and Claes H. de Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *Annals AAPS* 659: 122–131.
- Curran, James. 2009. "Narratives of Media History Revisited." In *Narrating Media History*, edited by Michael Bailey, 1–21. London: Routledge.
- Dahlgren, Peter. 1992. "Introduction." In *Journalism and Popular Culture*, edited by Peter Dahlgren and Colin Sparks, 1–23. London: Sage Publications.
- Deacon, David. 2007. "Yesterday's Papers and Today's Technology. Digital Newspaper Archives and "Push Button" Content Analysis." *European Journal of Communication* 22 (1): 5–25.

- Fink, Katherine, and Michael Schudson. 2014. The Rise of Contextual Journalism, 1950s–2000s." *Journalism* 15 (1): 3–20.
- Flaounas, Ilias, et al. 2014. "Research Methods in the Age of Digital Journalism." *Digital Journalism* 1 (1): 102–116.
- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.
- Grosan, Crina, and Ajith Abraham. 2011. *Intelligent Systems*. Berlin: Springer.
- Günther, Elisabeth, and Thorsten Quandt. 2016. "Word Counts and Topic Models." *Digital Journalism* 4 (1): 75–88.
- Handford, Michael. 2010. "What Can a Corpus Tell Us About Specialist Genres." In *The Routledge Handbook for Corpus Linguistics*, edited by Anne O' Keeffe and Michael McCarthy, 255–269. New York: Routledge.
- Harbers, Frank. 2014. "Between Personal Experience and Detached Information. The Development of Reporting and the Reportage in Great Britain, the Netherlands and France, 1880-2005." PhD diss., University of Groningen.
- Hartsock, John. 2000. *A History of American Literary Journalism. The Emergence of a Modern Narrative Form*. Amherst: University of Massachusetts Press.
- Jacobi, Carina, van Atteveldt Wouter, and Kasper Welbers. 2016. "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism* 4 (1): 89–106.
- Lee, Yong-Bae, and Sung H. Myaeng. 2002. "Text Genre Classification with Genre-Revealing and Subject-Revealing Features." *SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. New York: ACM. 145–150.
- Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. 2013. "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting & Electronic Media* 57 (1): 34–52.
- Manovich, Lev. 2012. "Trending: The Promises and the Challenges of Big Social Data." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 460–475. Minneapolis, MN: University of Minnesota Press.
- Manovich, Lev. 2015. "Data Science and Digital Art History." *International Journal for Digital Art History* 26 (1): 11–35.
- Mussell, James. 2008. "Ownership, Institutions, and Methodology." *Journal of Victorian Culture* 13 (1): 94–100.
- Nerone, John. 2010. "Genres of Journalism History." *The Communication Review* 13 (1): 15–26.
- Neuendorf, Kimberley A. 2002. *The Content Analysis Guidebook*. Thousand Oaks: Sage.
- Nicholson, Bob. 2013. "The Digital Turn." *Media History* 19 (1): 59–73.
- Riffe, Daniel, Stephen Lacy, and Frederick Fico. 2005. *Analyzing Media Messages. Using Quantitative Content Analysis in Research*. Mahwah: Lawrence Erlbaum Associates.
- Simon, A. F. 2001. "A Unified Method for Analyzing Media Framing." In *Communications in U.S. elections: New agendas*, edited by R. P. Hart and D. R. Shaw, 75–89. Lanham, MD: Rowman and Littlefield.
- Vella, Stephen. 2009. "Newspapers." In *Reading Primary Sources. The Interpretation of Texts from Nineteenth- and Twentieth-Century History*, edited by Miriam Dobson and Benjamin Ziemann, 192–208. London and New York: Routledge.

- Wijfjes, Huub. 2017. "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." *Tijdschrift voor Mediageschiedenis* 20 (1): 4–24.
- Yang, Tze-l., Andrew J. Torget, and Rada Mihalcea. 2011. "Topic Modeling on Historical Newspapers." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–104.
- Zheng, Zhaohui, Xiaoyun Wu, and Rohini Srihari. 2004. "Feature Selection for Text Categorization on Imbalanced Data." *ACM Sigkdd Explorations Newsletter* 6 (1): 80–89.

# IN SEARCH OF AMERICA

## Topic modelling nineteenth-century newspaper archives

**Quintus Van Galen and Bob Nicholson**

*This article considers how, and why, “Topic Modelling” tools can be used to analyse historical newspaper archives. While a growing number of media and communication studies projects have applied these techniques to corpuses of born-digital journalism, using the same tools to analyse large-scale collections of historical newspapers requires us to overcome additional technological and methodological challenges. Our discussion is framed around a historical case study examining references to the United States in the 19th Century British Library Newspaper Archive. The article begins by highlighting the problems that researchers of both digital and historical journalism face when attempting to deal with an enormous body of evidence. Next, it argues that Topic Modelling offers one potential solution to these problems by providing a way to “distant read” the archive. The remainder of the article is divided into five experiments that demonstrate how Topic Modelling can be applied to a series of research questions, each of which is applicable to other projects that might make use of newspaper archives. As well as demonstrating the investigative potential of topic modelling, the article also highlights the practical and technological barriers that currently undermine its effectiveness, particularly when it is applied to archives of historical material.*

### **Introduction**

In 2014, the British Library moved its collection of historical newspapers to a new, hi-tech storage facility at Boston Spa in the north of England. Some 60 million newspapers were transported from London by a fleet of lorries and are now stored in a low-oxygen warehouse that is operated remotely using robotic cranes. Each day, these curatorial robots glide through towering canyons of print, retrieving volumes of material from miles upon miles of shelving. The scale of this archive is staggering, and it grows each day as hundreds of newly published papers are ingested into the collection. A few of these publications are crumbling from overuse, but most have not been read since the day they were published. Many may never be touched by human hands again; historians will not summon them, and their robotic curators will forever pass them by in search of more popular titles. A similar story (but with fewer robots) might



be told about newspaper archives all over the world, and rapidly expanding collections of born-digital journalism are no exception; we could speak in similar tones about unrequested drive partitions and eternally idle server clusters where news data sits archived and unread. The overwhelming and ever-expanding size of journalistic collections is one of their most exciting qualities, but navigating this “vast *terra incognita* of print” (Leary 2004), in either its digital or physical forms, is a daunting prospect. There is, put simply, too much for us to read.

Scholars of historical and digital journalism are well-accustomed to this problem. In both fields, our research routinely requires us to accept the necessity of *not reading*; to find ways of deriving meaning from a vast body of texts that we can rarely explore in their entirety. Recent debates around the potential and challenges of working with big data (Lewis and Westlund 2015) have thrown this situation into sharp relief, but it is worth remembering that Victorian observers were as overwhelmed by the energy and expansion of their own print culture as we are now by the exponential growth and liquid nature of digital journalism (Fyfe 2008). For a human reader, the prospect of consuming 100,000 texts is hardly more practical than reading 100 billion of them; in this respect, scholars of journalism have been grappling for centuries with challenges that are conceptually similar to those posed by big data. However, the digital revolution has stimulated significant new advances in the techniques and technologies used to navigate these large-scale journalistic archives. As well as entrusting robots with the task of shelving our newspapers, an increasing number of researchers are now using different machines to interpret them.

This article considers the pitfalls and possibilities of a technique called Topic Modelling—a form of automated content analysis developed in the field of computer science. It begins by reflecting on the problematic ways in which we have traditionally coped with the unwieldy scale of our archives, before arguing that Topic Modelling offers one potential solution to these problems. After briefly explaining how the technique works, we use it to “distant read” a collection of nineteenth-century British newspapers. This article outlines five Topic Modelling experiments that we undertook in order to explore research questions that are common across many journalism history projects. In the case of this article, the experiments are structured around a specific historical case study that aims to unpack Victorian press coverage of the United States. What follows is by no means a comprehensive overview of all the ways in which Topic Modelling might be applied to the study, or indeed the practice, of journalism. Nor is it a triumphant celebration of this methodology; our experiments met with mixed success, and the historical insights we gained are partial, provisional and often unsurprising. Instead, we highlight the potential of this tool for the study of journalism history, whilst also outlining the formidable technical challenges faced by researchers who intend to apply digital humanities tools to archives of historical, rather than born-digital, newspapers. In the process, we highlight some of the shared challenges faced by scholars of digital and historical journalism in the age of big data, as well as how differences in the structure and content of our archives often impede the development of common tools, datasets and methodologies.

### *In Search of America*

In 1893, the eyes of the world fell on Chicago. That year, the Midwestern city played host to the World’s Columbian Exposition—a spectacular exhibition of art,

science and entertainment organised in order to commemorate the 400th anniversary of Columbus's arrival in the New World. Chicago was widely regarded as the shock city of the age; a young, restless and electrifyingly modern landscape that expressed America's growing confidence and promised its visitors an "early encounter with tomorrow" (Lewis 1997). Thousands of curious Europeans boarded steamships and made pilgrimages across the Atlantic in order to experience the "lightning city" at first hand. Millions of their compatriots, however, made the journey using a different technology—the printing press. As *The Spectator* put it, "the visit to Chicago which even the most stay-at-home portion of the world will make through its newspapers, will enable them to realise the great change that is coming over America" (Anonymous 1891). Sure enough, thousands of articles about Chicago and its World's Fair appeared in British newspapers and periodicals. Victorian readers could "visit" the city by perusing the columns of London papers such as *The Times* and the *Daily News*, provincial alternatives like the *Leeds Mercury* or the *Northern Echo*, illustrated weeklies like *The Graphic* and *The Sketch*, comic magazines like *Punch* and *Ally Sloper's Half Holiday*, periodicals such as the *Saturday Review* or *The Strand*, special-interest publications like the *Art Journal* or the *Musical Times*, and hundreds of other titles plucked from British news-stands. Most of these papers ran lengthy articles about Chicago, but readers would also have encountered passing references to the city in news stories, financial bulletins, joke columns, sports reports, advertisements, serialised fiction and countless other journalistic genres. As a result, most "stay-at-home" Victorians did not "visit" Chicago in a single sitting; their glimpses of the city were diffused across thousands of fragmented textual encounters.

All of which poses a problem for historians. How do we begin to make sense of a process that involved thousands of articles published across hundreds of newspapers? A full-text search for the word "Chicago" in the *Nineteenth Century British Library Newspapers* databases returns 54,342 results for the years 1892–1894, which is probably too many for even the most diligent historian to read and analyse in full. The scale of the problem becomes even more apparent when we widen the parameters of our search. The vicarious "visit" to Chicago made by British readers in the early 1890s was by no means unusual. In fact, as Nicholson has argued elsewhere (Nicholson 2012), these fragmentary encounters with America were an everyday occurrence for millions of Victorian newspaper readers and played a key role in mediating transatlantic relations during the nineteenth century. A full-text search for "America OR United States OR New York OR Chicago OR Yankee OR Dollar"—terms we have identified as the most common, and least ambiguous, markers of America—between the years 1860 and 1900 returns a staggering 2.8 million results.

This is not a new methodological problem, and nor is it limited to those who study the Victorian press. Dealing with the overwhelming scale of media archives is a challenge that transcends historical periods and shapes the research of most scholars who work on journalistic material. If we shifted the focus of our case study to examine contemporary British perceptions of the United States during the Trump administration, we would be faced with an equally daunting volume of material drawn from digital journalism, social media and other big datasets. Even projects that adopt a narrower focus—on a particular event, or an individual newspaper—will often find themselves dealing with thousands of potentially relevant articles. In the pre-digital era, researchers typically responded to this problem in two ways. Firstly, they focused their analysis on

a limited sample of newspapers. For instance, Troy Bickham's (2009) impressively broad-ranging study of how the British press responded to the American Revolution is based on a sample of newspapers and magazines that encompasses approximately a third of those published during the conflict. This is by no means an insubstantial body of evidence, but Bickham understandably acknowledges that "a complete reading of every issue would be too arduous a task" (11). His sample was carefully constructed in order to include papers from different political positions and geographical locations, but not all works of journalism studies achieve this balance. Researchers have often privileged a small number of canonical publications, usually a country's leading metropolitan dailies. *The Times*, for instance, looms particularly large in histories of nineteenth-century Britain, partly because the publication of Palmer's Index from 1867 onwards made it easier to locate relevant articles. As Andrew Hobbs (2013) argues, *The Times* was largely unrepresentative of wider Victorian journalism practices, and its "deleterious dominance" over our scholarship has too long obscured the importance of provincial and popular newspapers. Joel Weiner's study of the Americanization of the British Press (2011), for instance, focuses exclusively on London papers and thereby ignores the agency, inventiveness and influence of their regional rivals. Even when these issues are carefully considered, it is difficult to pinpoint a sample of newspapers that is both manageable and representative.

Even if a perfectly representative sample of newspapers can be determined, the volume of material in these titles usually prohibits exhaustive reading. A typically issue of *The Times* during the nineteenth-century could exceed 100,000 words. Add a few dozen other newspapers to the sample, and a historian will be faced with more than a million words to wade through for each new day of the period they have chosen to examine. Not even the best-intentioned researcher will read these publications from cover-to-cover. Instead, journalism historians typically combine sampling with selective reading. Sometimes we focus on particular sections of the newspaper, such as the leader or editorial pages. For other projects we zoom directly to particular dates that are likely to feature material on our chosen topic. Alternatively, we skim through a lengthier run of issues, allowing our eye to be drawn in by promising headlines. This approach can generate meaningful results, but it tends to focus our attention on eye-catching events rather than the banal (but nevertheless powerful) rhythms of everyday journalism. For a case study like ours, this "top down" approach (Nicholson 2013b) might reveal something useful about how leader-writers and foreign correspondents responded to Chicago during the months of the World's Fair, but there's a good chance that we would miss hundreds of passing references to the city in domestic news, advertisements, serialised fiction or joke columns. In short, both of these traditional ways of coping with scale—sampling and selective reading—typically result in a very partial analysis of the press and discover things chiefly in the places where we consciously decide, in advance, to look for them.

The digitisation of historical newspaper archives provides us with powerful new tools for dealing with their scale. The simplest and most widely used of these tools is undoubtedly the keyword search. This "bottom up" method (Nicholson 2013b) treats the digitised text as if it were an index to the source, directing the researcher's attention straight to the page and sentence that matches their search terms. In our case, it allows us to quickly locate references to America across a much wider range of

publications and journalistic genres than the non-digital techniques outlined above. A typical page of search results lists the *Manchester Times* and the *Musical Times* alongside their more illustrious London namesake; snippets from adverts and joke columns brush shoulders with leaders and editorials. The established canon of journalism history has already been profoundly destabilised by these digital archives, and many historians are now examining provincial newspapers and popular magazines which they would never have chosen to consult in print. Moreover, keyword search tools are allowing us to pursue research projects that were previously unimaginable (Nicholson 2013b). But it has important limitations. While a keyword search can locate 2.8 million references to America in a matter of seconds, it cannot analyse them—manually reading this quantity of evidence remains as impractical as ever (Huistra and Mellink 2016). Faced with hundreds of pages of search results, researchers often fall back on partial (and sometimes rather unsystematic) forms of close reading.

Consequently, an increasing number of researchers have begun to experiment with different methods of “distant reading”—a term first proposed by the literary critic Franco Moretti (2005). The most common implementation of this philosophy focuses on word frequencies. By counting the number of times that a word or phrase occurs within a newspaper archive, and tracking how this changes over time, we can reveal large-scale, *longe durée* insights into the language use of a historical corpus, or the shifting visibility of particular ideas, topics and individuals (Vliegenthart, Boomgaarden, and Boumans 2011; Nicholson 2012; Kestemont, Karsdorp, and During 2014; Lansdall-Welfare et al. 2017). These analyses are often (but not necessarily) visualised in the form of a line histogram, or n-gram, and as a result this is often colloquially referred to as the n-gram method, although it can also be found (in historiography) under umbrella term such as Cultural Analytics and Culturomics. For our case study, an n-gram search for “Chicago” in British newspapers reveals an initial spike in coverage in 1871 (when the city famously burned to the ground), a steady increase throughout the 1880s as the rapidly re-built metropolis accumulated economic and political power, and a clear peak in 1893 when the World’s Fair took place. As well as highlighting potentially significant historical moments, this quantitative search method offers a crude way to measure and compare the journalistic presence of particular words and, by extension, their associated subjects. In 1893, for example, the word “Chicago” appeared in approximately 5.5% of the newspaper articles featured in the British Library’s database. The term “New York,” by comparison, consistently appeared in 8–12% of articles between 1860 and 1900, which prompts us to consider the relative importance of these cities. However, these searches tell us very little about the *context* in which any of these words appeared. A passing reference to New York is weighted the same as a full-page editorial about Chicago; an article criticising the World’s Fair has the same impact on the graph as one that praises it. Moreover, the researcher needs to make *a priori* choices about words that unambiguously signify the concept under investigation. Does the word “Boston”, for instance, chart articles about the city in Massachusetts, or the town in Lincolnshire? Finally, without diligent close reading, it is easy to misinterpret spikes in the graph. A single advert, if repeated often enough, can significantly skew the results—as can a particularly active racehorse, if its name is also a keyword.

This article examines a different approach to “distant reading” digital archives. Topic Modelling potentially offers the best of both worlds, allowing both the distant reading of a large corpus, while retaining some of its contextual meaning. In Topic Modelling, each text, or document, is considered to be constructed out of a number of topics. These topics represent a set of words which share a significant pattern of co-occurrence (two or more words that are used together in one text) or cross-occurrence (words that are not necessarily used together in one text, but which share a third word with which they are). For example, “Alabama”, “claim” and “Washington” could be one topic, and “Gladstone”, “parliament” and “debated” another. The phrase “Alabama claims debated in parliament” would therefore be classified as containing both of these topics; 50% of the phrase comes from topic one (“Alabama” and “claims”) and 50% comes from topic two (“debated” and “parliament”). The word “in” belongs to a list of commonly occurring “stopwords” that we instruct the Topic Modelling algorithm to ignore. Crucially, these topics are not predetermined by historians, but are automatically identified by the Topic Modelling algorithm and presented to the researcher for identification. In other words, the algorithm discovers relationships between a set of words, but does not know what this relationship, or any of the words for that matter, *mean*. It is left to the historian to determine that the words “Gladstone”, “Parliament” and “Debated” signify a topic concerned with political news. The advantage of this approach is that—unlike keyword searches, n-grams, or what Boumans and Trilling (2016) refer to as “supervised methods” of automated content analysis—it requires no *a priori* assumptions from the researcher about the content of their archive. We instruct the Topic Modelling algorithm to find what *it* thinks the most prominent topics are, then use our historical expertise and interpretive skills to make sense of its discoveries.

Despite being developed in the early 2000s, Topic Models have not been broadly used in historical research. They have seen some use in language and literature studies, as tools to explore the evolution of a writer’s tone and style, or to classify diary entries (Blevins 2010; Erlin 2014). Other researchers have employed the method successfully on modern news articles, which can be scraped from the web as a ready-made corpus (Kawata and Fujiwara 2016). The earliest research on historical newspapers was performed by Newman and Block (2006), who applied topic modelling to 72 years of transcribed text from the *Pennsylvania Gazette* in order to better understand the corpus they were using for their historical research. Their work remains of great importance, as it proved Topic Modelling a historical newspaper was possible and could produce valid interpretations. More recently, Nelson (2010) explored the topics that could be found in the *Richmond Daily Dispatch*, in order to shed light on journalism in the confederate states throughout the American Civil war. He identified topics such as “Fugitive Slave Advertisements”, “Patriotic Poetry” and “Military Recruitment”, and tracked their evolution throughout the conflict. Topic modelling is now being used more extensively in modern journalism and communication studies. The studies employing them are as varied as the fields they represent: analyses of political speeches (Quinn et al. 2010); press coverage of nuclear issues (Jacobi, van Atteveldt, and Welbers 2016); the identification of influential news items (Krestel and Mehta 2008); Twitter commentary (Malik and Pfeffer 2016); general content analysis of a particular newspaper or periodical (Kawata and Fujiwara 2016; Kestemont, Karsdorp, and During 2014); or attempts to solve problems related to authorship attribution (Savoy 2013). Recent studies have

found topic models to be more versatile and better suited at classification tasks than other methods, such as those relying on a predefined dictionary (Guo et al. 2016). However, as was pointed out by Günther and Quandt (2016) in this journal, these uses rely on clean and accurate data, often sourced online and often “born digital”.

This paper builds on these foundations by exploring the viability of topic modelling a different kind of newspaper dataset, one that is both much more historical and much more dirty. Rather than focussing on a single paper, our project uses the entire corpus of the *19th Century British Library Newspaper* archive (parts 1 and 2). This dataset contains 72 different newspapers, or approximately 15 million articles. This is a substantial dataset, but it is important to stress that it covers only a small fraction of the total newspapers published at this time. At present, and indeed for the foreseeable future, the use of digital archives does not bypass our reliance on sampling. Titles in our archive were selected for digitisation by an academic advisory panel, who attempted to identify historically significant papers from a range of political and geographic backgrounds. While their choices and omissions are open to debate, they have nevertheless resulted in an archive that is *fairly* representative of Britain’s newspaper press at this time. Working with a dataset of this size potentially allows us to reach conclusions about the presence of America in “the Victorian press” as a whole, and to conduct comparative analyses between a range of different newspapers and locations. But this potential comes at a price. Both the *Pennsylvania Gazette* and *Richmond Daily Dispatch* were digitised through “re-keying” (manually transcribing the pages), which is extremely accurate and often achieves error rates of less than five percent. Our archive was digitised by optical character recognition [OCR] software, and is therefore significantly less accurate (Tanner, Muñoz, and Ros 2009; Hitchcock 2013). Moreover, during the digitisation process, each newspaper page in our dataset was segmented into its constituent articles. This automated process is not entirely reliable, which means that multiple articles are often joined together—particularly on pages of small advertisements. As a result, it is sometimes difficult for our Topic Modelling algorithm to determine when an article about America ends and the next part of the column (usually an article on an unrelated subject) begins; it treats both articles as a single, unified document. This means that normalisation of the data is nearly impossible, which causes problems for Topic Modelling (Günther and Quandt 2016). As the normalisation steps are supposed to ensure the documents that enter the modelling stage are as uniform in terms of OCR, segmentation and tokenisation as possible, the issues with segmentation could jeopardise the validity of the resultant model. For example, if advertising pages are not segmented into their individual components, then the model will regard an advert for imported American beef as being part of the same document as dozens of other adverts for unrelated products. It is impossible to fix these issues of segmentation and OCR entirely; both require significant technological breakthroughs, or an enormous amount of manpower to manually correct. However, this article demonstrates effective ways to work with a messy dataset and still derive meaningful results.

Keyword searching and n-grams have now become relatively accessible methods for exploring digital newspaper archives, largely because they have been integrated into many commercial user interfaces. However, Topic Modelling is more of a DIY pursuit and currently requires a higher degree of technical expertise, as well as an investment in the necessary hardware. In an ideal world, Topic Modelling would be

```

=====
topic no: 3
0.054*"steamer" + 0.038*"arrived" + 0.025*"left" + 0.025*"here" +
0.017*"york," + 0.016*"line" + 0.016*"general" + 0.015*"mail" +
0.013*"london," + 0.010*"company's" + 0.010*"april" + 0.010*"to-day."
# 0 top article ID: W01_GWHD_1882_05_17-0010-035    top article score:
0.995433|
mail and line steamer from london, has arrived here. may left here to-day
for may line steamer clan to left here to-day. from and trom leit here
to-day for i and port-said 15.-steamer from the left here to-day for may
from bombay, arrived here bombay, may left here for bombay, may 16.
eteamer from steamers from from from 11 liverpool; and from may 13.-
steamer from arrived here to-day may 16 steamship company's steamer has

```

### FIGURE 1

An extract from a typical Topic Model. This is the third topic in a 20-topic sequence of articles mentioning “America” between 1860 and 1899. The extract shows the “topic words”, information about the “top article” associated with them, and the first lines of the “top article”.

performed using the high-powered servers common in the field of Computer Science. But, as our project demonstrates, it is possible for a pair of humanities researchers to make progress with more modest equipment. The Topic Modelling algorithm we chose to employ is Latent Dirichlet Allocation (LDA), first proposed by Blei, Ng, and Jordan (2003). LDA is reasonably fast in its operation, capable of reading and modelling 10–15,000 words per minute. More importantly, LDA is also the most-used Topic Modelling algorithm in the digital humanities, thanks to its relatively simple concept and metrics, and its availability as packages and plugins for a wide variety of popular programming languages. This paper uses the Gensim package’s implementation of LDA (Rehurek and Sojka 2010); access to this package and all other functionality is handled through a Python script. The newspaper dataset, which we bought directly from its publishers Gale Cengage, sits on a 10TB NAS drive, which we access using an Elasticsearch host and a Raspberry Pi. The actual Topic Modelling took place on a modern desktop computer (i5-7400 processor, 16GB RAM). The power of this machine determines the length of time required to generate each topic model. Processing all of the 2.8 million articles mentioning America would take many weeks for each of our queries, and our basic hardware often struggled to handle this volume of data. We therefore limited ourselves to samples of up to 5000 randomly selected articles per year—the limit Elasticsearch can return in a single query. For experiments that were not tied to individual years, we used samples of up to 50,000 articles.

The initial results of a Topic Modelling algorithm are difficult to fully represent in a conventional academic article. In our case, the output takes the form of a text file containing a predetermined number of topics. Each topic contains a list of “topic-words” with their respective chance to signify the topic, and a list of the five best-matching articles. Figure 1 shows an example topic, which clearly features words that we might expect to find in news stories about the arrivals and departures of ships. In this case, the presence of the word “steamer” has a 5.4% chance of indicating this topic. The best-matching document (referred to here as the “top article”) fits the topic 99.5%. Once this list of topics has been generated, it is up to the researcher to begin interpreting them. Our workflow began with an examination of the “topic words” for a particular topic, followed by a close-reading of its “top articles”. We use this

**TABLE 1**

Three sample entries from an annotated 20-Topic model of articles mentioning “America” between 1860 and 1899.

Keyword	Years	Topic set	Words	Annotation
America	1860–1899	20	0.036**“prices” 0.031**“market” 0.028**“trade” 0.021**“good” 0.019**“6d.” 0.017**“business” 0.017**“supply” 0.016**“quiet” 0.016**“sold” 0.015**“lb.” 0.015**“foreign” 0.014**“rather”	Market News
America	1860–1899	20	0.029**“prince” 0.017**“princess” 0.012**“royal” 0.011**“spain” 0.011**“duke” 0.010**“revenue” 0.010**“emperor” 0.010**“chamber” 0.010**“de” 0.010**“duchess” 0.009**“wales” 0.009**“minister”	Royal News
America	1860–1899	20	0.091**“ar” 0.048**“” 0.032**“~” 0.026**“.” 0.023**“en” 0.020**“*” 0.018**“er” 0.016**“ol” 0.015**“th” 0.014**“ij” 0.012**“te”	OCR errors / Undetermined Topic.

information, alongside our existing historical expertise, to try and determine the subject of each topic. Some of these subjects—such as the shipping news mentioned above—are clear, while others are more difficult to interpret. In some cases, we were unable to successfully determine the nature of a topic. This is partially a result of our “dirty” dataset, but similar puzzles would emerge using even the most accurately digitised newspapers. After all, a linguistic pattern that seems significant to an algorithm may not be meaningful to a human reader. As part of this process, we create an “annotated” topic model (Table 1), from which we can draw preliminary conclusions about the dataset, or identify areas for closer reading. Alternatively, these may be visualised to better show the relative importance of the individual words to the topic’s meaning (Figure 2).

There are many ways in which press historians might make use of Topic Models. They can be used to explore the press as a whole, but can also be focused more





**FIGURE 2**

Relative importance of words within topics.

narrowly on particular publications, journalistic genres, historical periods or search terms. The following sections of this article describe a series of experiments that we conducted on our own dataset, with a view to exploring the presence of America in Victorian newspapers. However, the methods that we use could easily be applied to projects exploring other aspects of historical and digital journalism. In the first section, we model the press as a whole in order to determine whether America was a prominent topic of interest. Next, we focus our attention more closely on articles that specifically mention the United States. In this experiment, we begin by modelling the full-text of these articles, but also consider the benefits and drawbacks of topic modelling simply their headlines. Our third experiment considers how topic models can be focused, either by selecting specific newspapers or identifying historical periods and events of interest. Finally, our last experiment demonstrates how topic models can be used to measure and visualise large-scale changes in the press. It is important to stress that these experiments reveal the limitations of topic modelling as much they demonstrate its potential. Our intention here is to highlight methodologies that work, but also to document some of the blind alleys that we wandered down in the process.

### Experiment One: Attempting to Detect America

Firstly, we wanted to determine whether it is possible to identify a topic as “America” without making any *a priori* assumptions or selections. In other words, did the United States occupy a sufficiently prominent place in British newspapers for it to appear as a distinctive topic without us deliberately looking for it? To test this, we selected a semi-random sample of the archive: a total of 23,266 articles, with their headlines, from 26 titles. Once this sample had been isolated from the main dataset, we needed to determine the number of topics that we wanted our algorithm to generate. This is the key “human” choice in topic modelling, and requires careful consideration and experimentation. If the number is too small, we will be left with topics that are too broad to be meaningful; imagine, for instance, if we tried to divide everything in a newspaper into just two linguistic camps. A larger number of topics brings us more detail, but can also introduce problems. Generating 10,000 topics, for instance, would be too many to analyse and annotate manually, thereby negating the practical advantages of distant reading. It might also identify very narrowly focussed topics and prevent us from recognising broader historical patterns. A useful topic model strikes a balance between these two extremes. Unfortunately, there is no

hard rule for selecting the best number of topics. Instead, finding a value that works for a given research question or dataset is an iterative process that should be undertaken for each query. For our project, we tested the topic generation algorithm with four different numbers of topics: 10, 20, 40 and 100, in order to establish the right level for this dataset.

Our 20-topic search effectively identified some of the period's most prominent journalistic genres. Shipping news (characterised by words like "steamer" and "arrived"), classified advertisements ("wanted", "price", "street"), financial bulletins ("stock", "ditto", "limited"), theatrical notices ("messrs.", "musical", "christmas"), court reports ("prisoner", "defendant", "witness", "charged") all emerge clearly. None of these will come as a surprise to press historians, but there is value here in confirming something that we already know. Firstly, it suggests that we should pay attention to topic models when they uncover patterns that *do* surprise us. Secondly, we are able to determine the percentage of the corpus that falls into these predictable categories. This allows us to perform a kind of automated content analysis, akin to the work that used to be performed manually by measuring the column inches devoted to particular subjects or genres. This method could be productively applied to projects that seek to measure the changing nature of journalism, or to comparatively analyse the content of specific newspapers. This was not a focus of our particular case study, but we have demonstrated how such data could be visualised in our final experiment.

A 100-topic model reveals a series of more specific genres and subjects, including a topic that contains debates on women's suffrage ("franchise", "league", "women's", "voted"), one concerning the royal family ("royal", "prince", "majesty", "sir"), one reporting on "new patents" ("improvements", "manufacture", "apparatus", "lbs."), and another on North East England's iron industry ("Cleveland", "pig", "iron", "Middlesbrough"). Unfortunately, we did not find a topic uniquely defining "America" in our random sample at any level of detail. However, our 20-topic model did encounter the United States in two distinct places: theatre and entertainment adverts, and market news (particularly references to "dollars" or "dols."). In these cases, America was not a defining element of the topic, but it was frequently mentioned in the sample articles that were the topics' best representatives.

### **Experiment Two: Modelling Articles that Mention America**

Our first experiment revealed some of the contexts in which America was casually mentioned in the press, and that these casual encounters were at least prevalent enough to occur in the topic models of a random sample. However, we wished to go deeper and investigate how America was discussed in more detail. Therefore, we narrowed our corpus to articles where the word "America" or "American" appeared in the text. Our modest hardware could not handle all of these articles, so we selected a representative sample. This was done through a search query before we started modelling, which generated a corpus of 190,000 articles; approximately 4000 per year. We again experimented with different numbers of topics to obtain the most useful distribution. Once again, there is no magic number here; the patchy quality of our dataset means that some unintelligible topics are always introduced, no matter how many we pre-

select as our total. Eventually, we settled on a model composed of 40 topics, mirroring the work done by other researchers (Nelson 2010).

As expected, America features regularly in items that discuss transportation, such as the shipping timetables, and in the goods and stock market tables. However, we also found America deeply ingrained in the sports articles; both those aimed to inform gamblers and those reporting on scores. The role played by sporting journalism in shaping transatlantic relations in this period evidently needs to be explored in more detail by historians. Additionally, we found very distinct topics for the advertisements, both in the shape of personal ads reporting on marriages, births and deaths, and in the shape of mail-order schemes. In this latter case, America was often used as a byword for innovation and quality, particularly for fine mechanical devices. It is interesting though, that despite one producer proclaiming to "have followed the example of the Americans", they are still touted as "English watches, ... made by the most Eminent English Sculptors". The complex negotiation of national identities in this advert reaffirms the necessity of close reading—there are levels of nuance here that only a human reader can decode. America also seems particularly prevalent in adverts for various medicines and cures, to the point that in the slices for individual decades, we find topics constructed of words like "pill", "chemist", "cure" and "medicine"; clearly words descriptive of health-related advertisements. These cures are either presented as American inventions, or more commonly, described as being a large success across the Atlantic. In both of these cases, America is seemingly invoked as a seal of innovation and quality in medicine. As with sporting journalism, there is more work to be done on how these advertisements shaped British perceptions of American as a land of technological progress.

America's status as an emerging economic juggernaut is well represented, with three distinct types of economic articles appearing in our topics. Its presence is least felt on the goods market, where it seems to be incidentally referenced in the shape of a barrel of American apples or bushels of American wheat. Yet, in the other two categories, the Stock- and Money markets, the "Almighty Dollar" looms large. Daily papers regularly featured extensive updates on the rise and fall of stocks in American companies, and the fluctuations in the exchange rate between pound and dollar. The length and frequency of these bulletins increased significantly following the completion of the transatlantic telegraph cable in 1866. They provided practical intelligence for readers whose personal or commercial interests were tied to the American economy, but also spoke to a wider political context. As Nicholson (2013a) has explored elsewhere, economic competition with America was the cause of growing concern in Britain during the final decades of the nineteenth century. The presence of daily "American Markets" bulletins in the country's newspapers was one of the contexts in which this financial power was encountered. Even readers who preferred to focus their attention on columns of fashion advice and court reports would have noticed the words "American" and "New York" looming ever larger over the financial pages. Even if these columns sat at the periphery of many Victorians' reading experiences, we should never discount the power of something which sits constantly in the corner of the eye. Topic Modelling is particularly good at drawing our attention to material that seems unremarkable in isolation but assumes new significance when we recognise its pervasive and repetitive presence.

A close reading of the parliamentary pages, informed once again by our topic model, revealed that America was regularly discussed by British politicians. Here, too, there was a feeling that America was surpassing Britain when it came to economic and technological innovations. The *Birmingham Daily Post* reported in 1887:

[An MP] intends to submit a resolution in favour of the repeal of the Electric Lighting Act of 1882, which, owing to its impracticability, may be said to have destroyed our electric lighting industry, while in the United States of America, where there are no such restrictions, the same industry employs 50,000 workmen, and a capital of about £50,000,000. (Anonymous 1887).

This article could, of course, have been found using a conventional keyword search, but the topic model helps us to place it in a wider context, and connect it to a series of other articles that deployed similar language in response to the United States. We found evidence of transatlantic relations being shaped by the growing confidence of the United States as a nation, as seen in the Alabama claims and the refusal of American authorities to pursue Irish nationalists. The latter, in particular, was a sore point for the British government, which saw this as tacit support for an Irish rebellion. Once again, this is hardly a revelatory discovery; historians of transatlantic relations already know about the tensions sparked by Irish politics in both the nineteenth century and beyond. Nevertheless, the fact that this emerged as such a clear topic serves to demonstrate its scale and significance as a subject of debate. It also suggests that topic models will reveal meaningful insights when we apply them to research questions that are *not* already well understood. It is in this early, exploratory phase of a research project that these methods of distant reading may prove to be most useful, as Jacobi, van Atteveldt, and Welbers (2016) found in their work on modern newspapers.

### **Experiment Three: Focusing on Headlines**

Our attempt to topic model articles mentioning America produced promising results, but the efficacy of this method was continually undermined by the poor quality of OCR data and failed article segmentation. We initially attempted to address this problem by “slicing” our articles into smaller extracts, reasoning that in order to be found in the search, there had to be at least some clear text there. For each search result, we extracted a passage of text that began 100 words before the first mention of our keyword (“America”), and extended 500 words after its last mention. By doing this, we hoped to address some of the problems created by the poor segmentation of articles and avoid analysing sections of the newspaper page that were not related to America. This method produces results that are similar to the concordances used by corpus linguists, but in this case they extend much further in each direction from the central keyword. We hoped that this would reduce content from unrelated articles and eliminate large clusters of OCR errors, which tend to be grouped together in faded sections of a newspaper page where our keyword is unlikely to be discovered. However, despite the loss of large portions of this “noise”, we were surprised to discover no significant improvement in the quality of our topics. There is evidently much more work to be done on fine-tuning this “slicing” method, or on devising other automated techniques for dealing with bad OCR and poorly segmented articles.

Having failed to perfect our slicing method, we experimented with another way of bypassing bad data—focussing purely on an article’s headline. While it may seem counterintuitive to reduce a method designed to handle large bodies of text to short headlines, this approach has successfully been used to chart the common topics discussed in conference papers (Hall, Jurafsky, and Manning 2008). By feeding these headlines into a topic model, we hoped to map how the contexts in which the country was discussed were categorised by newspaper editors themselves. Unfortunately, many Victorian newspapers did not adopt headlines until the later part of the century, or gathered multiple items under a single column-header such as “Foreign News”. Moreover, when the archive was digitised, archivists responded to the absence of conventional headlines by entering the first line of an article as its “headline”. On the plus side, these articles were all manually corrected, which means the data quality is good. We found numerous topics of interest that were clearly defined, although many were obviously from articles where America was not the main focus. When we increased the number of topics modelled to 100, we saw very specific, clear groupings. Interesting topics that appeared were travel reports from MP’s to their constituents, articles about railways, the southern states during the civil war and the debates surrounding free trade. One particularly interesting topic consists of articles giving British readers a view of the everyday lives of their transatlantic cousins. These included articles on the American manufacturing industry, American elections, and American taxes and tariffs, typically written by Americans or foreign correspondents for the consumption of the British public. Each of these topics can now be earmarked for close reading, safe in the knowledge that they represented a statistically significant portion of America’s overall presence in Victorian newspapers.

#### **Experiment Four: Narrowing the Corpus**

The preceding experiments all focused on a wide range of newspapers, with a view to determining broad patterns in how America was covered by the Victorian press. However, we would not expect every newspaper to engage with the country in the same way; a paper’s politics, geography and format all shaped the contexts in which it looked across the Atlantic. Moreover, we would expect this coverage to change over time in response to political, economic, social, cultural and technological developments. All of these potential differences are obscured by an approach that seeks to speak for “the press” as a whole. As such, our next experiment focussed on two ways to narrow our corpus. First, we elected to focus our attention on specific newspapers. Topic modelling has been employed successfully on such data before, most notably by Newman and Block (2006) and Nelson (2010). However, their data were re-keyed, and therefore much cleaner. We selected two papers for the 20-year period from 1880 to 1899: the *Pall Mall Gazette* and the *Birmingham Daily Post*. The former was chosen because its editor for the first half of this period, W.T. Stead, was famously fascinated by America, and we might therefore expect the country to feature heavily in its reporting. The latter paper was chosen as a “typical” daily provincial newspaper that provides insights into the day-to-day portrayal of America in the British press. As an added technical benefit, the *Birmingham Daily Post* has been digitised to a comparatively high standard of OCR, leaving us with cleaner text than average.

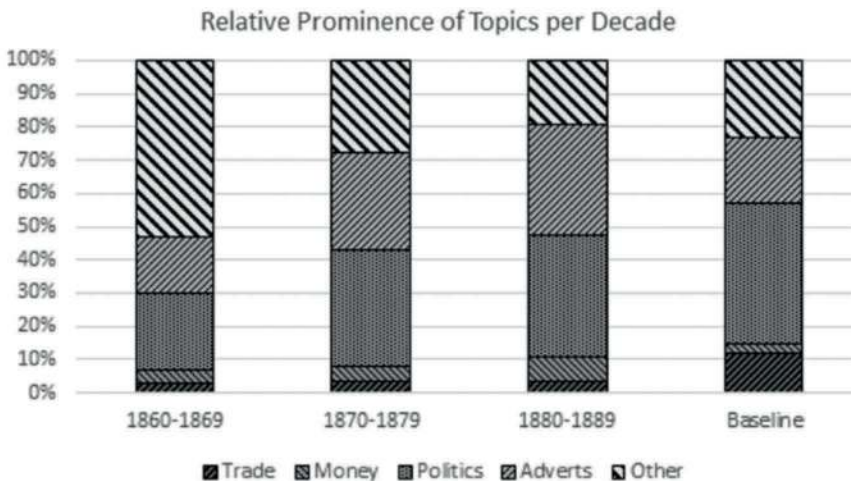
The *Birmingham Daily Post* mentioned America, or one of its other signifiers like "United States", in 16,365 articles during the last two decades of the nineteenth century. Our topic model managed to recover some of the many places in which a "Brummie" could expect to encounter his transatlantic cousins and competitors. One set of articles that emerges distinctly at both 20 and 40 topics modelled is a regular section titled "The American Iron Market", which featured reprints from a Wisconsin newspaper named *The Ironminer*. This particular aspect of the American economy was predictably of great interest to industrialists in the area since Birmingham was a metal-working centre. Additionally, we observed the presence of America in the monthly book reviews that the paper produced based on copies sent to their offices. Most of these were not of books published in America, or written by American authors, but had America as their subject matter. The *Pall Mall Gazette's* also displayed a clear topic for its books column, and, like the Birmingham paper, regularly featured American authors and publications. Once again, we also identified a recurring topic representing America as a source of innovation in fields like electric cooking and colour photography (Anonymous 1894; H.C.M 1898). Finally, the combination of a smaller sample size and clean OCR also allowed us to generate more meaningful topic models. For example, one collection contained articles intended to provoke panic or outrage, with titles like "Death in a snowdrift", "The Massacre of the Missionaries" or "The Perilous State of the Atlantic". Our enduring problems with article segmentation mean that not all of these stories have America as their main subject. Nevertheless, it shows the opportunities that topic models offer for determining the editorial tone of a paper at any given time.

Next, we investigated the use of topic modelling for gaining insights into press coverage surrounding a known historical event. To test this, we returned to the Chicago World's Fair and produced a narrower search focused on the period 1891–1894. While we did identify one topic made up primarily of articles about the Exposition, it was also widely discussed in the topics of political news, criminal news and business bulletins. These are precisely the kind of topics that might have been missed by conventional search and sampling methods, but which evidently shaped British encounters with, and responses to, the city. Future modelling on this question would probably benefit greatly from the clearer text generated by using a more advanced tokenisation and stemming package. Tokenisation is the process of dividing a piece of text into words: instinctual for humans, but sometimes tricky for computers. For these experiments, we defined a word as "the sequence of characters between spaces", which left hyphens in place, making "exhibition" and "ex-hibition" appear as two different words to the LDA algorithm. Stemming addresses a similar problem with variance in word forms: it ensures that the algorithm considers "exhibit", "exhibition" and "exhibited" the same word for modelling purposes. Again, these are the things that humans will understand without needing to be told, but computers fail to do. In our current programme, all of the words that a human would consider identical can end up in different topics. LDA does not "know" any language; it only traces patterns of word use, without understanding the words. Unlike research on modern sources, where these issues are much easier dealt with during data retrieval and normalisation and can often be reduced to trivial levels (Günther and Quandt 2016), they remain much larger stumbling blocks when looking at older material.

### Experiment Five: Change Over Time

Now that we are able to identify a topic's character, what do we use it for? One possibility is a comparative approach, modelling a variety of slices and seeing how topics within each slice change over time. We found the most convenient way to do this was to calculate the relative percentages of a topic within the overall corpus. Since how well an article matches a topic is expressed as a percentage, by averaging these percentages, taking into account the variance in article length, we can estimate how much of the corpus is devoted to that topic. Looking at the changes in these percentages over decades allows us to consider a temporal dimension to the data. For the three decades 1860, 1870 and 1880, using the 50,000 articles in which America was mentioned, we modelled 20 topics and calculated the percentage scores for each. We organised them into five broad categories: "trade", "money", "politics", "adverts" and "other". We did the same for our semi-random corpus of 23,000 articles, to serve as a baseline comparison with the press as a whole (Figure 3). Visually represented like this, the model can actually aid us in discovering interesting trends in the corpus; the temptation with research that relies on numerical data is to produce visualised numbers that look good, but add little to our understanding of history.

Not only is there is a significant evolution over time, the "America" topics are also present in different proportions to the baseline sample. Immediately obvious is the growth of adverts, from 17% in the 1860s to 33% in the 1880s. This might be understood as not only a reflection of the ever-growing cost of running a newspaper in the late nineteenth century, but also the presentation of America as a land of commerce and technology. Moreover, the decline of tariffs and the increase in transport links enabled the British market to be opened up to American companies. Also notable is the steady percentage of trade reports, which hardly changes over 30 years, hovering around 3%. The sharp decline of the "Other" category in the 1870s is also telling, although not for our understanding of the past as much as our understanding of the archive. It represents improvements in OCR quality between the two decades, facili-



**FIGURE 3**

Relative prominence of identified topics per decade.

tated by better preservation and higher-quality printing, and therefore higher identification of specific topics. This broad representation of America's shifting presence in the British press hints at the possibilities for using topic models as a form of automated content analysis. Such methods would benefit enormously from higher quality data—perhaps obtained from archives of more modern newspapers—and the identification of more specific topics. Our chart, for instance, is too broad to detect the growing popularity of imported American humour columns in British newspapers during the 1880s. Once again, we run into the limitations of working with a large but “dirty” dataset.

## Conclusions

British visitors to the Chicago World's Fair often observed that the city offered an intriguing glimpse into the future; an “early encounter with tomorrow” that hinted at how life in the “Old World” might soon be electrified and accelerated by new ideas and technologies. But they were also quick to point out the problems inherent in these new ways of living, as well as the stumbling blocks that would undermine their most utopian applications. Our experiments with Topic Modelling arrive at a similarly mixed conclusion.

On the downside, researchers hoping to employ Topic Modelling on large-scale digitised archives of historical newspapers need to be aware of the problems caused by poor data quality. Garbled OCR and erratic article segmentation confuse the Topic Modelling programme, because it assumes that all words in an article are correct and that they relate to the same subject. These errors also pose problems for researchers, who are initially prompted to close-read articles in the form of error-strewn transcriptions, rather than images of the original newspaper page. We believe that this is the major factor in understanding why our topic models struggled to pick up fine details akin to the results achieved by earlier work. Therefore, an improvement in data quality is needed in order for Topic Modelling to reach its fullest potential. The digitisation projects that generated the data we are using for our research ran between 2003 and 2009, which means that while the scanning software used was modern at the time, compared to 2018 standards it produced text of substandard accuracy. One solution, therefore, would be to re-OCR the entire dataset using new technology. Even this, however, will not remove errors entirely. This prompts us to consider ways of managing poor OCR, rather than waiting indefinitely for a solution that comprehensively fixes everything. Our attempts at “slicing” articles into relevant extracts were not initially successful, but the principle behind them deserves further investigation. We also attempted to spell-check articles before loading them into the Topic Modelling software. This produced more promising results, but increased computation time by such a factor that we were unable to pursue it using our relatively modest equipment. It would be beneficial for all researchers working on specific historical datasets to combine their resources and focus on ways to clean up these archives and prepare them for analysis. Unfortunately, access to the *19th Century British Library Newspapers* corpus is currently restricted by complex copyright agreements, making this kind of collaborative work difficult. We are likely to make more progress on newspapers held in open access archives such as *Chronicling America*, *Trove* and *Delpher*.



However, the problems caused by dirty data need not be fatal. As our experimental case study demonstrates, it is possible to derive meaningful results from these flawed archives. When used correctly, and in conjunction with a closer reading of the corpus, Topic Modelling provides a valuable new tool for researching the history of the press. It helps us to analyse an unreadably large corpus of journalistic texts in order to identify patterns and make a more informed selection for close reading. In turn, this empowers us to consult a much wider range of newspapers and magazines than were commonly manageable using pre-digital methods of sampling and selective reading.

Crucially, Topic Modelling keeps the context of a keyword's use intact during the modelling and analysis by looking at the document in which it occurs as a whole, rather than at the word alone, or a limited concordance surrounding it. Additionally, Topic Modelling does not rely on prior assumptions by the researcher. This latter point is particularly useful if the objective of a study is exploratory, and the goal is to establish an overall idea of what a newspaper, or an archive, contains. In our case study, we were able to tentatively map the shifting presence of America across dozens of nineteenth-century British newspapers. The topic models we generated confirmed many of our pre-existing assumptions about the subject, but also directed us to new and unexpected areas for close reading. They also prompted us to reassess the comparative significance of journalistic genres such as adverts and financial bulletins, and to measure how this changed over time. It will take more than these topic models to fully explain the relationship between the United States and the British press, but they provide a valuable birds-eye-view of a previously unfathomable volume of material.

Finally, on a broader level, our experiments with Topic Modelling highlight some of the connections and fissures that currently exist between the fields of historical and digital journalism studies. Researchers working on the nineteenth- and the twenty-first-century media are united by the challenges of dealing with scale; both must find ways to explore, interpret and represent bodies of evidence that are too expansive to read in their entirety. In this respect, both fields have much to gain from the collaborative development of Topic Modelling and other forms of automated content analysis. But the application of these tools is heavily dependent on the structure and quality of journalistic datasets, which are shaped in turn by the contemporary and historical challenges of archiving different forms of news media. The liquid and exponentially expanding landscape of online journalism is extremely challenging to capture, store and analyse (Karlsson and Sjøvaag 2016; Widholm 2016), but its born-digital text and metadata make it well-suited to the application of large-scale computerised analysis. While archives of historical journalism are often more stable and centralised, converting centuries of material into a digital format is an expensive and time-consuming task. Moreover, this digitisation process only partially bridges the gap between the worlds of print and digital journalism. It successfully converts paper and ink into a computer-readable format, but the faltering accuracy of this process prevents the resulting data from achieving full parity with born-digital content. The different archival practices and challenges faced by both fields means that collections of digital and historical journalism are rarely integrated and usually require the development of different tools and research methods. This can impede collaboration between scholars of digital and historical media and will be particularly problematic for comparative and *longue-durée* projects that use digital tools to trace the development of journalistic practices and

cultural trends across this divide. As the study of digital journalism increasingly becomes part of the practice of journalism history, finding ways to bridge these gaps between our divergent datasets will be important. While we are unlikely to achieve a seamless integration of historical and born-digital journalism, the experiments outlined in this article demonstrate that it *is* possible to apply similar forms of automated content analysis to both. It will be a long time before all of the newspapers in the British Library's archive are scanned and rendered accessible to these digital tools, but the development of techniques like Topic Modelling suggest that, one day, a time may come when none of them need to be left unread.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

## ORCID

Bob Nicholson <http://orcid.org/0000-0002-0863-963X>

## REFERENCES

- Anonymous. 1887. "Parliament, Yesterday." *Birmingham Daily Post*, July 7.
- Anonymous. 1891. "The Chicago Exhibition." *The Spectator*, July 18.
- Anonymous. 1894. "Cooking by Electricity!" *Pall Mall Gazette*, April 2.
- Bickham, Troy. 2009. *Making Headlines: The American Revolution as Seen Through the British Press*. DeKalb: Northern Illinois University Press.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.
- Blevins, Cameron. 2010. "Topic Modeling Martha Ballard's Diary." <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>
- Boumans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit." *Digital Journalism* 4 (1): 8–23.
- Erlin, Matt. 2014. "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864." *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. 91–114.
- Fyfe, Paul. 2008. "The Random Selection of Victorian New Media." *Victorian Periodicals Review* 42 (1): 1–23.
- Günther, Elisabeth, and Thorsten Quandt. 2016. "Word Counts and Topic Models." *Digital Journalism* 4 (1): 75–88.
- Guo, Lei, Chris J. Vargo, Zixuan Pan, Weicong Ding, and Prakash Ishwar. 2016. "Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling." *Journalism & Mass Communication Quarterly* 93 (2): 332–359.
- Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. "Studying the History of Ideas Using Topic Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 363–371.

- H.C.M. 1898. "Colour Photography: What Really Has Been Done." *Pall Mall Gazette*, September 21.
- Hitchcock, Tim. 2013. "Confronting the Digital: Or How Academic History Writing Lost the Plot." *Cultural and Social History* 10 (1): 9–23.
- Hobbs, Andrew. 2013. "The Deleterious Dominance of the Times in Nineteenth-Century Scholarship." *Journal of Victorian Culture* 18 (4): 472–497.
- Huistra, Hieke, and Bram Mellink. 2016. "Phrasing History: Selecting Sources in Digital Repositories." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49 (4): 220–229.
- Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. 2016. "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism* 4 (1): 89–106.
- Karlsson, Michael and Helle Sjøvaag. 2016. "Content Analysis and Online News." *Digital Journalism* 4 (1): 177–192.
- Kawata, Shinya, and Yoshi Fujiwara. 2016. "Constructing of Network from Topics and Their Temporal Change in the Nikkei Newspaper Articles." *Evolutionary and Institutional Economics Review* 13 (2): 423–436.
- Kestemont, Mike, Folgert Karsdorp, and Marten During. 2014. "Mining the Twentieth Century's History from the Time Magazine Corpus." *EACL* 2014. 62.
- Krestel, Ralf, and Bhaskar Mehta. 2008. "Predicting News Story Importance Using Language Features." *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.
- Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences* 114 (4): E457–E465.
- Leary, Patrick. 2004. "Victorian Studies in the Digital Age." In *The Victorians Since 1901: Histories, Representations and Revisions*, edited by Miles Taylor and Michael Wolf, 201–214. Manchester: Manchester University Press.
- Lewis, Arnold. 1997. *An Early Encounter with Tomorrow: Europeans, Chicago's Loop, and the World Columbian Exhibition*. Chicago: University of Illinois Press.
- Lewis, Seth C. and Oscar Westlund. 2015. "Big Data and Journalism: Epistemology, Expertise, Economics and Ethics." *Digital Journalism* 3 (3): 447–466.
- Malik, Momin M., and Jürgen Pfeffer. 2016. "A Macroscopic Analysis of News Content in Twitter." *Digital Journalism* 4 (8): 955–979.
- Moretti, Franco. 2005. *Distant Reading*. 2nd ed. Verso Books.
- Nelson, Robert K. 2010. "Mining the Dispatch." <http://dsl.richmond.edu/dispatch/pages/intro>
- Newman, David J., and Sharon Block. 2006. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society for Information Science and Technology* 57 (6): 753–767.
- Nicholson, Bob. 2012. "Looming Large: America and the Late-Victorian Press, 1862–1902." PhD Dissertation, University of Manchester.
- Nicholson, Bob. 2013a. "The Old World and the New: Negotiating Past, Present, and Future in Anglo-American Humour." In *History and Humour: British and American Perspectives*, edited by Barbara Korte, 151–170. Bielefeld: Transcript Verlag.
- Nicholson, Bob. 2013b. "The Digital Turn." *Media History* 19 (1): 59–73.

- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228.
- Rehurek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*: 45–50.
- Savoy, Jacques. 2013. "Authorship Attribution Based on a Probabilistic Topic Model." *Information Processing & Management* 49 (1): 341–354.
- Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros. 2009. "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive." *D-Lib Magazine* 15 (7/8)
- Vliegthart, Rens, Hago G. Boomgaarden and Jelle W. Boumans. 2011. "Changes in Political News Coverage: Personalisation, Conflict and Negativity in British and Dutch Newspapers." In *Challenging the Primacy of Politics*, edited by Kees Brants and Karin Voltmer, 92–110. London, UK: Palgrave Macmillan.
- Widholm, Andreas. 2016. "Tracing Online News in Motion." *Digital Journalism* 4 (1): 24–40.
- Wiener, Joel H. 2011. *The Americanization of the British Press, 1830s–1914: Speed in the Age of Transatlantic Journalism*. Basingstoke: Palgrave.

# JOURNALISM HISTORY, WEB ARCHIVES, AND NEW METHODS FOR UNDERSTANDING THE EVOLUTION OF DIGITAL JOURNALISM

**Matthew S. Weber** and **Philip M. Napoli**

*Archived webpages are a critical source of data for understanding the current state of the news media industry, as well as how the industry has changed over time. Dramatic changes in the news media industry in recent decades have occurred in tandem with the evolution on the Web. Archived webpages are valuable records for understanding and analyzing how newspaper companies have adapted to technological changes such as social media feeds and sharing of news content via Twitter. This article outlines a methodological approach to utilizing Web archives as a means of examining change in the news media industry. Researchers have developed new tools to improve researcher access to archived Web data in order to advance studies of the Web, and to enable the tracking of changes in news media as they emerge over time. A case study examining local news in the United States is used to illustrate the methodological challenges and promise of working with these data, highlighting the power and potential of Web archives for journalism research. Finally, the closing sections discuss challenges associated with the scale and scope of archived Web data and point to new areas for future research.*

It has been said that news content is the “first draft of history,” and yet we are quickly losing the first draft of our online history as digital preservation efforts struggle to capture online news media (Rosenzweig 2003). This situation is not new. Libraries, archives and other memory institutions have struggled in the past to capture printed news content. However, the end result has generally been more successful than has been the case in the realm of electronic media. As electronic media such as radio and television emerged as important sources of news, the challenges of—and gaps in—news archiving became much more pronounced (see Napoli and Karaganis 2007). In response, media law scholar Lawrence Lessig (2004) asked, “why is it that the part of our culture that is recorded in newspapers remains perpetually accessible, while the part that is recorded on videotape is not? How is it that we’ve created a world where

researchers trying to understand the effect of media on nineteenth-century America will have an easier time than researchers trying to understand the effect of media on twentieth-century America?" (p. 111).

The challenges of effective and robust news archiving are amplified in the era of big data, social media content, and interactive online news packages (Hansen and Paul 2015). Technical, legal, financial, and logistical challenges abound, including the growth of news applications (via mobile phones), interactivity and database connectivity, permission rights to preserve content created by third parties, and software architecture, among others (Boss and Broussard 2017; see, also, their article in this volume). With news outlets entering and exiting from the digital journalism landscape with high frequency, the volatility impacts the ability to preserve websites. Low barriers to entry and challenges to monetization exacerbate this issue (Radcliffe and Ali 2017), and create challenges for digital preservation. Nevertheless, online newspaper archives provide a critical record of social activity (Brügger and Schroeder 2017).

In response to extant challenges associated with identifying and preserving news media on the Web, this article has a dual focus: (1) this article discusses the role of Web archives in preserving newspaper content, and specifically focuses on developing an approach to archiving local newspaper content, and; (2) this article presents a case study of local news archiving in order to demonstrate the feasibility of this approach. Methodological challenges are discussed, as are potential analytical approaches in using these data. The balance of a methodological discussion and case study provide a roadmap for scholarship in this domain. In sum, this article sets a forward-looking agenda for leveraging the power of Web archives to better understand both the history and the future of today's ever-changing news media landscape.

### **Defining Archived Web Data**

Archival news media have been used in prior research to understand the transmission of news articles across different media platforms (Leskovec, Kleinberg, and Faloutsos 2007), to recreate hyperlinking patterns of news media organizations online, and to assess changes over time (Weber 2012; Weber and Monge 2014) and to explore social movements and collective action (Bennett 2005). In the digital humanities, scholars have established the importance of Web archives such as news media repositories as important artifacts for understanding the way in which culture was reflected locally at a given point in time (Gomes and Costa 2014). Archived webpages provide humanists with a rich history of society at a single point in time (Milligan 2016), and given the role of technology in modern society this point cannot be understated.

In order to better understand the practice of Web archiving, it is important to distinguish between digitization and Web archiving. Digitization of printed newspapers refers to the process of taking printed newspapers and translating the printed version into a searchable digital record. In prior generations researchers often had to travel to the newspaper itself, or to a particular library, in order to access hard copy records or microfiche. The move to digitization of printed newspaper content opens new avenues of research (Nicholson 2013; Bingham 2010), and many Web archives do provide access to digitized content. Web archiving is a separate practice from digitization. Web archiving is the preserving and archiving of versions of webpages so that they can be

recreated on a variety of platforms, regardless of how standards change over time. Web archiving is particularly important for newspaper content because the content changes frequently, and old stories are often lost as new versions and new iterations are created. And, of course, as traditional printed newspapers decline in number, to be replaced (to some extent) by digital-only outlets, a growing proportion of the journalism being produced exists exclusively online. In sum, Web archiving practices help to preserve digital-only newspapers, traditional online newspapers as well as other digital repositories. Some have gone as far as to dub digital news archives as the archeological survey site for the twenty-first and twenty-second centuries (Gaff 2017).

### **Challenges of Archived Web Data**

Archived Web data provides a unique opportunity to view newspaper content as it has existed online, in some cases dating from present date back to the mid-1990s and the early days of the World Wide Web. Key challenges associated with the creation of Web archives include the completeness of Web archives, challenges associated with dynamic data, problems stemming from changes in programming languages and ultimately, the cost of preserving content.

#### *Completeness*

As Masanès (2006) notes, most Web archiving today is either site-, topic-, or domain-centric. It is nearly impossible to crawl and store an entire copy of the Web, and therefore librarians, archivists and researchers have to make decisions about how to focus crawling efforts in order to serve a given set of aims. These choices often involve deciding to focus on a specific website or set of websites, crawling, and archiving based on a given topic or set of keywords, or crawling and archiving based on a domain (e.g. crawling and archiving the Australian Web domain—.au—as is done by the National Library of Australia and the Internet Archive). In the context of newspapers and news media, the rise of paywalls creates an increasingly common barrier to Web crawling (Ayala 2016). Paywalls create barriers that cannot be automatically crawled. Often the presence of paywalls requires that a site be manually crawled, or that researchers take the added step of purchasing access and gaining access outside of the actual Web crawling process.

Beyond user-based decisions to crawl specific domains or subsets of the Web, technological barriers increasingly create challenges for Web crawling. For instance, the evolving nature of Web advertising, including the growth of embedded videos, dynamic graphics, and interstitial ads that pop up in the transition between webpages, means that much of the advertising content present on a given webpage cannot be archived or replayed after the fact (Jessen 2010). The sprawl of Web content also means that there are challenges associated with identifying the appropriate news sources beyond national outlets (e.g. identifying local news outlets, tracking niche media outlets; Nielsen and Schröder 2014). Recent endeavors use iterative processes of crawling and re-crawling to improve completeness of the record of a single webpage, but completeness across an entire domain continues to lack (Jones and Neubert 2017).

### *Dynamic Data and Databases*

As the Web has evolved, websites have become increasingly dynamic and dependent on server-side resources such as databases. Social media websites, for instance, are constantly changing as users add content and post more frequently. Review websites, such as Yelp.com, allow users to search keywords and terms, and to return queries from dynamic databases. But Web archiving technology does not allow archivists to capture the richness and interactivity of these websites. More often than not, Web archiving efforts will capture a snapshot of these types of webpages, but much of the rich content contained on these websites will be lost unless the content is preserved by the owner of the website. As a result, archivists will often capture portions of these webpages, but users will be unable to access any of the dynamic elements of a given webpage because that portion of the webpage is not archived.

### *Changes in Programming Languages*

The programming languages underlying websites have evolved rapidly over the past two decades. Web archiving requires an archivist to capture the content on a webpage, but also to capture enough of the underlying code to be able to recreate or regenerate the webpage. New ways of displaying content, such as HTML5, mobile applications, JavaScript and Flash, create challenges for archivists as it is increasingly difficult to capture the code associated with these programming languages and platforms (Dougherty and Meyer 2014). Files created by leading organizations and start-ups in the newspaper industry range across the spectrum of file-types (.html, .xml, .php, .pdf). Because there is no standard practice for producing or managing news content within this evolving space, it is hard to develop a standard approach to archiving. Often, Web archivists simply do the best that they can to capture pages containing new programming languages, but the result is that some webpages cannot be retrieved or accessed later.

### *Cost of Web Archiving*

There are significant costs associated with Web archiving, from the initial work of determining what to archive, to the more technical costs associated with storage and maintenance of data over time (Dougherty and Meyer 2014). Archives are often plagued by the high cost of maintaining data; such challenges should not be underestimated as they can ultimately be the downfall of archiving efforts (Zimmer 2015). Preservationists often must take on costs of maintaining storage space and providing access, with little hope of recuperating expenses.

As a result of many of these challenges, once a Web archive has been created it is often difficult to retroactively locate desired webpages. Preservation does not guarantee that researchers will be able to access the necessary content, and often large chunks of Web-based content can be lost for good (Anat 2016).

## **Methodological Approaches**

Awareness of the limitations of Web archiving is important, particularly in the case of newspaper content, as understanding of the limits helps the researcher to



understand the pros and cons of working with this type of data. When turning to Web archives as resources, researchers have choices with regards to the data that they use. It is not always necessary to create new archives; there are many substantial Web archiving efforts currently underway, and decision to use existing archives as opposed to creating new archives is important.

Broadly, a number of ongoing research initiatives are aimed at improving research access to archival data sources. The Archives Unleashed initiative aims to create user friendly programs and tutorials to help social scientists and digital humanities scholars to work with Web archives (see <http://archivesunleashed.org>). The BUDDAH Project in the United Kingdom sought to demonstrate the feasibility of using large scale Web archives to examine text and language (see <https://buddah.projects.history.ac.uk>). Additionally, there are currently at least 68 major Web archiving initiatives led by national archives, national libraries and private companies, capturing more than 584 billion files (Costa, Gomes, and Silva 2017).

### *Use of Existing Datasets*

The emerging field of Web historiography argues for the use of archived Web data as a key means of studying and examining recent history of key events online (Brügger 2012). In the humanities and social sciences, scholars are quickly recognizing the key role that Web archives occupy as a means of studying recent phenomena occurring in the online space (Gomes and Costa 2014). In many cases, existing archives provide a resource for conducting research. For instance, in the United States, the Library of Congress's Chronicling America project provides access to both archived physical newspaper pages, and directories of online archived newspapers, including data back to 1789 (as a archive available on the Web providing access to digitized content). The British Library has partnered with findmypast to digitize up to 40 million newspaper pages, although the project charges for full text content (<https://www.britishnewspaperarchive.co.uk>). Google News offers a searchable archive of hundreds of newspapers across a wide range of time periods and across a broad geographic space. Many organizations also offer their own Web archives, including newspapers such as *The New York Times*, *The Guardian (UK)*, and *The Times of India*.

Alternatively, broad Web archiving efforts also offer access to records of newspaper content. For example, Internet Archive ([archive.org](http://archive.org)) is one of the best-known Web archives in the world, and contains one of the largest repositories of archived newspaper content (in addition to their rich record of the World Wide Web at large). Using large-scale repositories such as the Internet Archive can allow for easier access to data, but the use of such repositories means that the researcher is reliant on others with regards to decisions about what is crawled, when it is crawled, and how much is retained from a single Web domain. Indeed, whenever a research is using a preexisting collection it is hard to know how the collection was developed, how decisions were made with regards to what to include, and what was excluded from the dataset. On the other hand, these archives are readily accessible with lower barriers to entry than custom archiving.

### *Creating New Web Archives*

It is also possible for one to create a new archive to suit a researcher's needs. Indeed, despite the presence of extant Web archives, digital preservation is not the norm for newspapers (Hansen and Paul 2015). A 2012 national survey by Educopia showed that most U.S. newspaper respondents maintain digital news records for fewer than 5 years, without ensuring the longevity of such records. A 2014 national survey by the Reynolds Journalism Institute of 70 digital-only and 406 hybrid (digital/print) newspapers echoed this, finding only 12% of hybrid newspapers reported backing up digital news content and 20% of digital-only newspapers report backing up none (Carner, McCain, and Zarndt 2014). Thus, regardless of the robustness of select newspaper archives, and the overall growth of Web archiving practices, there are major gaps with regards to the archiving of newspaper content. The known gaps in the archiving of newspaper content point to one of the many reasons why one might want to curate a new Web archive of news content for research purposes. Custom archiving of content allows archivists and researchers to control the content that is collected, to set a specific scope, and to ensure the accuracy, completeness, and regularity of data collection efforts.

There are many ways in which one can go about creating a new Web archive. A number of existing standards exist that provide guidelines for creating Web archives. For instance, the Web ARChive (WARC) file format is an ISO standard for preserving Web content. For-profit organizations such as Hanzo will help organizations to preserve their digital content. The Internet Archive has developed the Archive-IT platform, which makes Internet Archive archiving technology available to partner organizations for a fee.

Further, the preceding discussion illustrates how Web archives are a critical—but often incomplete—record of the ongoing transformation of news media. A key challenge, however, becomes the decision of what to archive. The following outlines a methodological approach to collecting and archiving news media websites.

### *Creating New Web Archives: The Case of Local News*

The methodological case study used hereafter is based on our experience creating an archive of a sample of local news websites for 100 communities in the United States. This research was undertaken as part of the study with the goal of understanding the health and robustness of local community news and the role that community characteristics (demographic, geographic, etc.) might play in affecting the robustness of local news (see, e.g. Napoli et al. 2017). The project aimed to create a national snapshot of local community news in 2017; the project was intended to create a framework for continuing to monitor changes in online local community news coverage. In order to create a bounded sample that (a) captured the breadth of newspaper websites across the country and (b) simultaneously captured the depth of newspaper coverage within each individual community as completely as possible.

Our methodological approach focuses on four key steps: (1) accurately defining the scope of the Web archiving project; (2) verifying the completeness of Web crawling within the defined scope; (3) determining the depth of the Web crawl necessary to capture sufficient information, and; (4) the development of an appropriate sampling

scheme to appropriately capture the breadth of coverage. Each of these steps are detailed in the following.

### *Scope*

First, it is important to recognize that it is not possible to capture everything on the Web. Every Web archive has its limits, and clearly defining what is and is not part of an archive is a key part of a successful project (Heil and Jin 2017). Scoping can be specific to Web content, but may also relate to the specifics of the research that is being conducted or the information that is being sought. In the case of this study, we bounded our data collection by limiting data to communities with a population between 20,000 and 300,000. Using US Census data, this provided a list of 493 communities. Moreover, we did not intend to collect each and every community, nor each and every website, within those 493 communities. Rather, we selected a random sample of 100 communities, and then verified that the sample approximated the population based on population, income, demographic composition.

### *Completeness*

The second key step in our methodological approach focused on verifying the completeness of our Web crawling efforts. The challenge of completeness is particularly acute in the context of news and local news; it is difficult to build an accurate list of local news sources, and many niche outlets are difficult to locate via traditional Web search (Nielsen and Schröder 2014; Tewksbury 2005). Our project relied on a partnership with the Internet Archive; we used the Archive-IT community archiving platform to create our own Web archive. The Archive-IT platform allows users to specify the websites that they would like to collect, and to set parameters including the frequency and depth of crawling (e.g. the number of links crawled away from the original page). There is a cost associated with the platform, but there are many open source tools available that provide comparable options.

Regardless of the platform for archiving, it is necessary to determine a set of websites to archive, and to determine how complete an archive one wishes to create. In our case, with a list of 100 communities our goal was to create as complete a sample as possible. The concept of community is central to what is local, and a full discussion of the way in which community is utilized in this work can be found in Napoli et al.'s (2017) examination related to this topic. Broadly speaking, rather than focusing on specific realizations of what a community is as defined by those living within the community, we chose to define community based on geographic boundaries.

The focus of this research was on local news sources; thus, we included local newspapers, as well as local radio stations, local television stations and local online only news sources. Our inventory of local news sources was limited to those sources geographically located within each sampled community; and thus excluded news sources that might produce news of relevance to the community, but that were geographically located outside of the community. This inventory was created through a systematic process of consulting multiple media databases and directories—11 in total (for a complete list see Appendix A) and supplementing these database and directory

scans with a multi-stage online keyword search protocol (for a description of the scanning and keyword search protocol see Appendix B). This multi-pronged approach reflects the fact that a comprehensive portrait of the news sources serving local communities today can only be achieved via cobbling together information from a broad array of sources. Even large-scale commercial media directories (e.g., Cision) were discovered to have coverage gaps when compared to the manual multi-database/directory search process described here

In the case of this research, the outlined query process generated a final list of 733 local news sources. Interestingly, we encountered very few instances in which a print or broadcast news outlet serving an individual community did not have an accompanying online presence, suggesting that, at this point, virtually all traditional media also make their news content available online—even the kind of small, local news outlets that characterized this sample. Thus, Web archiving would seem to have potential to gain insights into the entirety of a local media ecosystem.

It is also important to note that the rise of digital paywalls (even at the local level) provided an impediment to systematic archiving. As previously noted, Web archivists face challenges with regards to automatic archiving of webpages that are behind paywalls. In the case of the New Measures Research Project, we purchased accounts for each of the websites that had a paywall to enable crawling of the websites.

### *Depth*

Once the completeness of a Web archive has been determined, the next step is to determine the depth of a crawl. For instance, *The Dover Post* provides local news coverage of Dover, Delaware. The website for *The Dover Post* ([www.doverpost.com](http://www.doverpost.com)) has a series of subsections and is updated with relative frequency. Given the limitations of our funding and our research, it was not possible to capture every webpage on each of the 733 local news sources, nor did we have the resources to analyze all that data. Therefore, we decided to focus on the front-page coverage of each news source, and crawled to a depth of one. That means that in addition to archiving the front page, we archived each webpage that was one hyperlink (or one click) from the front page. For *The Dover Post*, that meant capturing the full content of all stories mentioned on the home page of that outlet. This approach was premised on the notion (well-established in the journalism studies literature) that news source front pages/home pages represent a meaningful indicator of the most important news events and issues affecting a community, and follows in prior traditions of sampling for content analysis of newspapers (for a discussion see Napoli et al. 2017).

### *Sampling*

The last methodological question in creating a Web archive is sampling. Again, working with limited resources and limited time, it was necessary to restrict the sampling of content. As a result, we decided to create a constructed week sample, selected a Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday at random over a 2-month period during the summer of 2016. The superiority of constructed week sampling compared to other approaches, such as continuous week sampling, is

well established in media studies (Riffe, Aust, and Lacy 1993). Thus, websites were crawled on July 27, August 2, August 6, August 28, September 9, September 21, and September 26. In the ideal, less time-constrained, data gathering scenario, such a constructed week sample would be drawn from the entirety of a calendar year, to better control for any idiosyncrasies in news occurrences or reporting that might be associated with narrower time periods. And, in the ideal scenario, a 2-week constructed week sample would likely be preferable as well. Given constraints associated with this research (time and funding) the above provided as complete a as collection as possible; moreover, prior work has established that a narrower time frame of constructed week sampling can be used in the case of online media without sacrificing the significance of the results (Connolly-Ahern, Ahern, and Bortree 2009). Finally, consideration was given to the time of day of a crawl. Due to technical limitations associated with the large number of webpages we were crawling, it was not possible to align the time of day each crawl occurred. Conversely, prior research examining constructed week sampling and time of day has found that the time of day of data collection does not significantly impact results with regards to online news media (Tanner and Friedman 2011).

### *Summary of the News Measures Archive*

Using this methodology, we created a robust record of the front-page coverage of local news in 100 communities. The Web crawl across the sample week generated a collection containing 1.6 million documents (html files, pdfs, images, audio files, etc.) and 2.2 terabytes of total data based on the seed set of local outlets (split across print, television, radio, and online-only news outlets). A process of manual evaluation of the front pages of the archived local news sources found that the archive contains just over 20,000 distinct news stories.<sup>1</sup>

### **Analytical Approaches**

The study of Web archives of news content affords the analysis of a variety of research questions. The use of archived Web data in research is relatively new, and as a result, associated analytic techniques are emergent. Researchers are working to develop new platforms for analyzing archived Web data, but many of these are still efforts in progress.

Extant methods provide a starting point for analyzing data. For instance, analysis of the content of webpages provides a mechanism for understanding how discourse of news from a particular newspaper, or on a particular topic, has changed over time (Matthew and Beatríz 2007). In the case of this study, we utilized content analysis of individual stories to determine the relationship between community geographic and demographic characteristics and the robustness of the local journalism. Each story was analyzed in terms of whether it is local, original, and addresses one of seven critical information needs; the seven critical information needs are emergencies and risks, health, education, transportation systems, environment and planning, economic development, civic information, and political life (Friedland et al. 2012). This content analysis was done in order to reach some generalizable conclusions about the types of

communities most in need of philanthropic, advocacy, or policy interventions on behalf of local journalism, and to explore the possibility that *journalism divides* affect communities in ways that parallel the well-known digital divide (see Napoli et al. 2017).

Future efforts could focus on other units of analysis. For instance, outlet-focused analysis could be employed to determine the types of outlets that make the most (and least) significant contributions to a community's journalism ecosystem. Outlet ownership represents another level of analysis that could be employed, to explore questions about the relationship between ownership concentration and journalistic output across communities and/or over time. And, of course, a focus on individual stories opens up a wealth of possibilities for looking at the nature and evolution of local journalism output (the content analysis that we have applied to these stories thus far represents only the tip of the iceberg in terms of analytical possibilities). The corpus of story data also provides a starting point for developing automated approaches to content analysis.

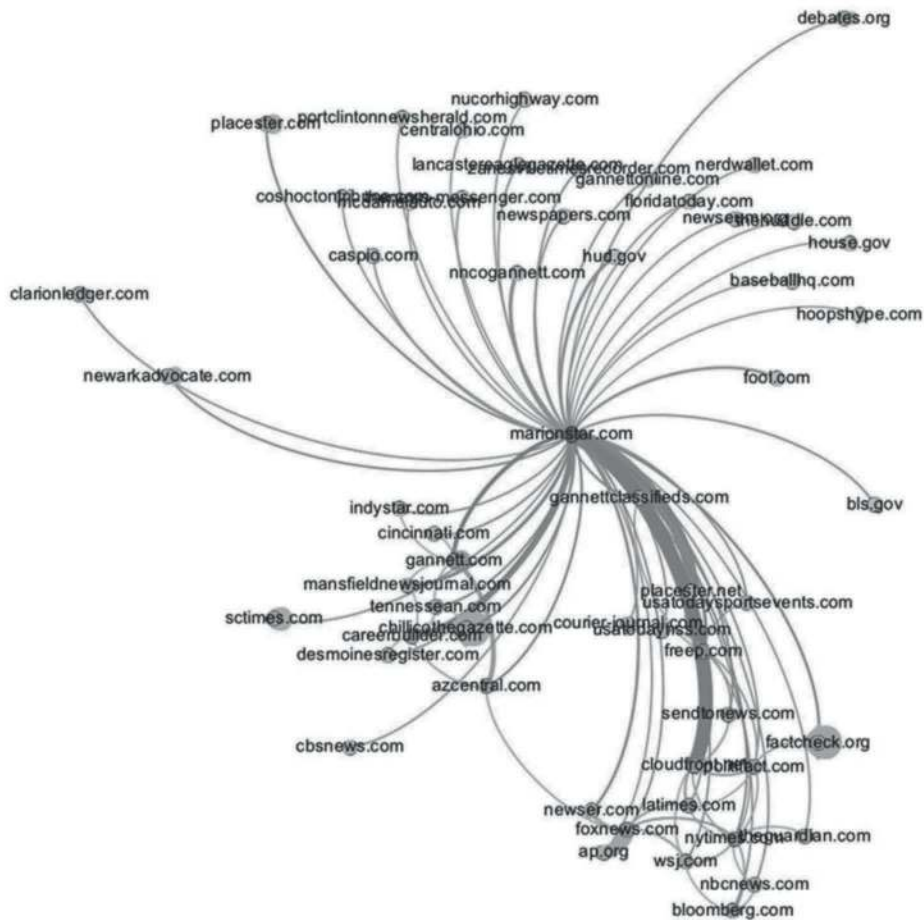
Aggregation of data allows for the summarization of topics and the analysis of key themes; similarly, aggregation provides a means of conducting network analysis, using connections between newspaper articles and newspapers to better understand the flow of information (David 2007). Recent work using Web archives demonstrates the utility of using social network analysis as a means of understanding the flow of news content between newspaper websites (Weber 2012, 2014). Additionally, entity extraction of key terms and locations provides a way to summarize large swaths of data and to understand trends in archived newspaper content (Martin, Pfeffer, and Carley 2013). The potential of these approaches is described in greater detail below.

### *Analyzing Networks in Local News with Archives*

In order to demonstrate the plausibility of analysis of Web archives, the following demonstrates the combination of both social network analysis and entity extraction as a means of understanding patterns that emerge in archived Web content. This is one approach among many, but was at the core of research questions in our study that focused on the locations discussed in published content and the localness of local news content.

One broad use of social network analysis is for general inspection of and description of the connections that exist between various actors in a network. For example, as illustrated in Figure 1, network analysis provides a means for visualizing the connections between websites in a data set and better understanding information flow. The visualization in Figure 1 shows the egonetwork of The Marion Star, in Marion, Ohio. This visual analysis shows how different newspaper organizations are connected to the Marion Star based on hyperlinks between two newspaper companies.

The visualization in Figure 1 was generated by using an API produced by the Internet Archive's Systems Interoperability and Collaborative Development for Web Archiving (WASAPI) project (<https://archive.org/details/wasapi>). The WASAPI project provides researchers with a suite of programming interfaces (APIs) that allows researchers to retrieve data in formats that are much easier to use for research. For instance, the WASAPI APIs allow scholars to generate link lists in a social network analysis format. Link lists can then be used with existing social network analysis packages (including Gephi and R) to quickly generate network visualizations and to analyze websites in a dataset. The full code for the visualization generation process using the API can be



**FIGURE 1**

Visualization of the Ego Network of The Marion Star (2016). Ego network including and websites that were connected to The Marion Star by hyperlinks that occurred more than three times in the given dataset. A connection exists between two websites if there was a hyperlink connecting the two websites. A threshold of three has been established in prior research as a critical threshold for determining a meaningful connection between websites based on hyperlinking activity within a given year (Weber and Monge 2011)

found at on GitHub<sup>2</sup>; this website includes a set of documentation that details the implementation of the API so that others can use the data to create similar visualizations. The visualization shown in Figure 1 was generated by importing data from the API into Gephi, The ForceAtlas2 algorithm was used for visualization generation in Gephi; this particular algorithm is designed to create spatial separation and allow for qualitative descriptive analysis of the network (Jacomy et al. 2014).

Thus, the visualization shown in Figure 1 demonstrates one way to analyze and visualize the data from a Web archive in order to understand key patterns of connection between newspapers, as well as to understand the behavior of a single newspaper. For instance, the bottom left quadrant of the visualization shows the connection between The Marion Star and other newspapers in the USA Today (Gannett) Network of newspaper. In the bottom right is a network of connection to national newspaper

outlets. The aggregate shows the distinct news sources used by a local news outlet. For instance, the visualization reveals hyperlinks pointing to the Commission of Presidential Debates' website (debates.org) and the website for the U.S. House of Representatives (house.gov), suggesting the importance of political reporting during the period of analysis. Because of the visualization filters out hyperlinks occurring less than three times during the constructed week, only key sources referenced frequently are shown. Finally, the visualization also points to some of the underlying technology supporting the website, such as caspio.com, which provides backend technical services for databases and applications. This approach could be expanded and used to compare the diversity of different local newspaper outlets by coding the types of connections and comparing the diversity of links between outlets.

### *Analyzing Place in Local News with Archives*

Another feature of the WASAPI APIs is the ability to generate named entities on a given website. The named entity feature reads the text that is stored on an archived webpage, and uses the Stanford Named Entity Recognition (NER) package to retrieve named city and state combinations lists in the Web archive record (Finkel, Grenager, and Manning 2005). Sanford's NER has been shown to have a relatively high degree of reliability in recognizing named place combinations in text documents (Ratinov and Roth 2009). For instance, returning to The Dover Post example, the NER feature allows the researcher to see every place that is mentioned in a local news story about crime in Dover, Delaware.

Returning to the prior illustrative example, and aggregating the list of places mentioned by The Dover Post, it is then possible to approximate how much of the coverage in The Dover Post focuses on Dover, Delaware, on Delaware more broadly, or on areas outside of Delaware. Replicating this out, it is then possible to look at the full range of news stories, and to see the places mentioned in those news stories. The API provides a means for converting to latitude and longitude, and thus calculating the difference between a given community and the places covered.

A test analysis of three days from this study, using 124,000 pairs of community and places mentioned, revealed an average local news distance of 542 miles (SD =225 miles). A visualization in Figure 2 shows a color coded map of 3 specific communities as a means of demonstrating the dramatic spread in local news coverage. It is worth noting that there is a wide degree of variation in the "locality" or "localness" of news for the communities that were sampled. The results are nevertheless substantial; much extant research on local news talks about local as being bounded within a given community. For instance, research looking at audience perceptions of what is local (Heider, McCombs, and Poindexter 2005), codifying local to examine the nature of content reported in a community (Gilliam Jr, Valentino, and Beckmann 2002), or looking at local news production (Hood 2007), has all focused on the community as the unit of analysis without considering the localness of the content being published.

This study points to an *average* distance that exceeds far beyond the boundaries of most towns and counties. The largest county in the United States is San Bernardino County in California; that county is 20,105 square miles, or approximately 142 miles wide. Our average distance stretches across nearly four such counties. In other terms,





and early fall. As previously noted in the text, the time of year likely does not impact the volume of content collected, but the timing with regards to the 2016 U.S. Presidential election likely skewed the topic of some of the coverage further towards political events than would normally occur. Moreover, the decision to collect only front pages limits the range of content collected in this dataset. For instance, sports news is content that may not always be featured on the front page, but is always present within a newspaper (Bridges 1989). While these issues do not seem to impact the findings of the present study, questions of timing and scope clearly have the potential to impact the results of analysis.

More broadly, the use of Web archives raises a number of broader questions that are important for consideration when conducting research in this space. First, the creation of Web archives involves collecting data created by other organizations; because these data are being copied and stored, there are important ethical and legal considerations that should be taken into mind. For example, copyright laws vary by country, and in some countries this type of Web archiving could be called into question. As others have suggested (Niu 2012), when it comes to copyright protections local laws should always be taken into consideration and it is best to contact legal counsel to determine what are acceptable best practices.

Moreover, there are ethical considerations to keep in mind as well. For instance, Web archiving of a newspaper's webpage may include the archiving of comments made by readers on a webpage. Comments often include identifying information, including a person's real name, and may result in identifying information being collected without permission. Again, national laws, such as those in the European Union pertaining to the right to be forgotten, may impact the ability to collect and reuse these types of data, and these concerns should be addressed at the time of data collection (for more on this, see boyd and Crawford (2012) for a thorough examination of key issues pertaining to the ethics of research using large-scale online data). Finally, the aggregate scope of data is also worth considering. In the example provided in this study, 2.2 terabytes of data were collected, which provided a sizeable testbed to explore issues of locality. It is, however, not always necessary nor pertinent to collect data on this vast a scale. Following on points made by boyd and Crawford (2012), it is important determine the type of data necessary in the context of research questions being asked. Big datasets are attractive for many reasons, but often create more of a headache than they are worth in terms of both cost to access and store, and time to clean and analyze.

### *Future Research*

Throughout the course of this discussion, a number of key questions have been raised at the intersection of Web archiving and news that merit future scholarship. First, this work points to an important discussion of what is defined as local news. Through the use of Web archives, it is also possible to monitor the changing nature of what is local using the approaches described in this research. To that end, using the framework of coding for critical information needs, and also analyzing the source of a given news article, it is also possible to more thoroughly examine the composition of local news within given communities, and to use Web archives to assess the diversity

of information in a community. Beyond questions of content, the archiving of news and local news websites also raises important questions about the technology underlying these websites. As online news is increasingly provided via social media and mobile platforms, new approaches to archiving are needed. For instance, mobile platforms provide a high degree of personalization (Thurman and Schifferes 2012; Westlund 2014). In response, it may be important to better understand how news—both local and national—change content for individual’s needs. This could be accomplished through a more thorough analysis of archived news websites but would require greater depth of crawling. Moreover, this work may enhance knowledge pertaining to the production of local news, as others have demonstrated the production processes are changing within newsrooms hand-in-hand with changes in the technology—and this is reflected in the way content is distributed to consumers (Widholm 2016).

### *Conclusion*

Web archives are the stewards of our cultural heritage (Dougherty and Meyer 2014), and provide an incredible resource for understanding how newspaper content and coverage has changed over time. They can be used to track the impact of media policy changes (e.g., changes in ownership regulation), philanthropic interventions, or changes in market or competitive conditions. Web archives also face notable challenges with regards to data collection, completeness and aggregation.

This article has outlined many of the key challenges and limitations associated with Web archiving today. At the same time, this article presents one approach for archiving local news content, and subsequently developing new research out of a given archive. Using social network analysis and named entity extraction, the combination of a method for data collection and an approach to analysis provide a framework for working with Web archives at scale to better understand the history and trends of news online.

Moreover, the discussion of the distance of local news presents a first-of-its-kind quantification of what constitutes local news. Big data analysis, in conjunction with Web archives, has the potential to open up critical new paths for research examining the evolving nature of news ecosystems, and potentially gives researchers and policy-makers an important tool for understanding how news is distributed around us. We see Web archives as a mechanism for understanding changes in news media over time. Web archives provide a direct lens for examining content as it was produced over time, and for addressing a wide range of critical research questions.

### **NOTES**

1. The complete archive is available at <https://archive-it.org/collections/7520>. Note that the actual archive today is larger than described in the text. The archive was created with specific boundaries for the purposes of this research, but the Internet Archive decided to continue the data collection effort on a monthly basis to maintain the record of local news websites.
2. <https://github.com/mwe400/LocationMapper>

## REFERENCES

- Anat, Ben-David. 2016. "What does the Web Remember of its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain." *New Media & Society* 18 (7):1103–19. doi: 10.1177/1461444816643790.
- Ayala, Brenda Reyes. 2016. "Challenges for Web Archivists: Issues in the Preservation of Digital Cultural Heritage." In *Annual Review of Cultural Heritage Informatics*, edited by Jennifer Weil Arns, 151–64. London: Rowman and Littlefield.
- Bennett, W Lance. 2005. "Social Movements Beyond Borders: Understanding Two Eras of Transnational Activism." *Transnational Protest and Global Activism*, 203–26. New York: Rowman and Littlefield.
- Bingham, Adrian. 2010. "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." *Twentieth Century British History* 21 (2):225–31. doi: 10.1093/tcbh/hwq007.
- Boss, Katherine, and Meredith Broussard. 2017. "Challenges of Archiving and Preserving Born-Digital News Applications." *IFLA Journal* 43 (2):150–7. doi: 10.1177/0340035216686355.
- Boyd, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15 (5):662–79. doi: 10.1080/1369118X.2012.678878.
- Bridges, Janet A. 1989. "News Use on the Front Pages of the American Daily." *Journalism Quarterly* 66 (2):332–7.
- Brügger, Niels. 2012. "Web Historiography and Internet Studies: Challenges and Perspectives." *New Media & Society* 15 (5):752–64. doi: 10.1177/1461444812462852.
- Brügger, Niels, and Ralph Schroeder. 2017. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press.
- Carner, Dorothy, Edward McCain, and Frederick Zarndt. 2014. "Missing Links: The Digital News Preservation Discontinuity." *Reynolds Journalism Institute*. Columbia: University of Missouri.
- Connolly-Ahern, Colleen, Lee A Ahern, and Denise Sevick Bortree. 2009. "The Effectiveness of Stratified Constructed Week Sampling for Content Analysis of Electronic News Source Archives: AP Newswire, Business Wire, and PR Newswire." *Journalism & Mass Communication Quarterly* 86 (4):862–83.
- Costa, Miguel, Daniel Gomes, and Mário J. Silva. 2017. "The Evolution of Web Archiving." *International Journal on Digital Libraries* 18 (3):191–205. doi: 10.1007/s00799-016-0171-9.
- David, Deacon. 2007. "Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis." *European Journal of Communication* 22 (1):5–25. doi: 10.1177/0267323107073743.
- Dougherty, Meghan, and Eric T. Meyer. 2014. "Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs." *Journal of the Association for Information Science and Technology* 65 (11):209–2195. doi: 10.1002/asi.23099.
- Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." Paper presented at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

- Friedland, Lewis, Philip Napoli, Katherine Ognyanova, Carola Weil, and Ernest J Wilson III. 2012. "Review of the Literature Regarding Critical Information Needs of the American Public." Communication Policy Research Network (CPRN). Federal Communications Commission. [https://transition.fcc.gov/bureaus/ocbo/Final\\_Literature\\_Review.pdf](https://transition.fcc.gov/bureaus/ocbo/Final_Literature_Review.pdf)
- Gaff, Donald H. 2017. "Extra! Extra! Read All About It: Newspaper Archives as Archaeological Site Survey." *Journal of Archaeological Method and Theory* 24 (2):451–65. doi: 10.1007/s10816-016-9273-3.
- Gilliam Jr, Franklin D, Nicholas A. Valentino, and Matthew N. Beckmann. 2002. "Where You Live and What You Watch: The Impact of Racial Proximity and Local Television News on Attitudes about Race and Crime." *Political Research Quarterly* 55 (4):755–80.
- Gomes, Daniel, and Miguel Costa. 2014. "The Importance of Web Archives for Humanities." *International Journal of Humanities and Arts Computing* 8 (1):106–23. doi: 10.3366/ijhac.2014.0122.
- Hansen, Kathleen A., and Nora Paul. 2015. "Newspaper Archives Reveal Major Gaps in Digital Age." *Newspaper Research Journal* 36 (3):290–8. doi: 10.1177/0739532915600745.
- Heider, Don, Maxwell McCombs, and Paula M Poindexter. 2005. "What the Public Expects of Local News: Views on Public and Traditional Journalism." *Journalism & Mass Communication Quarterly* 82 (4):952–67.
- Heil, Jeremy M., and Shan Jin. 2017. "Preserving Seeds of Knowledge: A Web Archiving Case Study." *Information Management* 51 (3):20–4.
- Hood, Lee. 2007. "Radio Reverb: The Impact of "Local" News Reimported to Its Own Community." *Journal of Broadcasting & Electronic Media* 51 (1):1–19.
- Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, A Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PLoS One* 9 (6):e98679.
- Jessen, Iben Bredahl. 2010. "The Aesthetics of Web Advertising: Methodological Implications for the Study of Genre Development." In *Web History*, edited by Niels Bruggen, 257–78. New York, NY: Peter Lang.
- Jones, Gina M, and Michael Neubert. 2017. "Using RSS to Improve Web Harvest Results for News Web Sites." *Journal of Western Archives* 8 (2):3.
- Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. 2007. "Graph Evolution: Densification and Shrinking Diameters." *ACM Transactions on Knowledge Discovery from Data* 1 (1):1–42. doi: 10.1145/1217299.1217301.
- Lessig, Lawrence. 2004. *Free Culture: How Big Media Uses Technology and Law to Lock Down Culture and Control Creativity*. New York: Penguin Press.
- Martin, Michael K., Juergen Pfeffer, and Kathleen M. Carley. 2013. "Network Text Analysis of Conceptual Overlap in Interviews, Newspaper Articles and Keywords." *Social Network Analysis and Mining* 3 (4):1165–77. doi: 10.1007/s13278-013-0129-5.
- Masanès, Julien. 2006. "Web Archiving: Issues and Methods." In *Web Archiving*, 1–53. Berlin: Springer.
- Matthew, Reason, and García Beatriz. 2007. "Approaches to the Newspaper Archive: Content Analysis and Press Coverage of Glasgow's Year of Culture." *Media, Culture & Society* 29 (2):304–31. doi: 10.1177/0163443707074261.
- Milligan, Ian. 2016. "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives." *International Journal of Humanities and Arts Computing* 10 (1):78–94. doi: 10.3366/ijhac.2016.0161.

- Napoli, Philip M., and Joe Karaganis. 2007. "Toward a Federal Data Agenda for Communications Policymaking." *CommLaw Conspectus: Journal of Communications Law and Policy* 16:53–96.
- Napoli, Philip M., Sarah Stonbely, Kathleen McCollough, and Bryce Renninger. 2017. "Local Journalism and the Information Needs of Local Communities." *Journalism Practice* 11 (4):373–95. doi: 10.1080/17512786.2016.1146625.
- Nicholson, Bob. 2013. "The Digital Turn." *Media History* 19 (1):59–73. doi: 10.1080/13688804.2012.752963.
- Nielsen, Rasmus Kleis, and Kim Christian Schrøder. 2014. "The Relative Importance of Social Media for Accessing, Finding, and Engaging with News." *Digital Journalism* 2 (4):472–89. doi: 10.1080/21670811.2013.872420.
- Niu, Jinfang. 2012. "An Overview of Web Archiving." *D-Lib Magazine* 18 (3/4).
- Radcliffe, Damian, and Christopher Ali. 2017. "Local News in a Digital World: Small-Market Newspapers in the Digital Age." *Tow Center for Digital Journalism*. New York: Tow Center for Digital Journalism.
- Ratinov, Lev, and Dan Roth. 2009. "Design Challenges and Misconceptions in Named Entity Recognition." In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 147–55. Boulder, Colorado: Association for Computational Linguistics.
- Riffe, Daniel, Charles F Aust, and Stephen R Lacy. 1993. "The Effectiveness of Random, Consecutive Day and Constructed Week Sampling in Newspaper Content Analysis." *Journalism Quarterly* 70 (1):133–9.
- Rosenzweig, Roy. 2003. "Scarcity or Abundance? Preserving the Past in a Digital Era." *The American Historical Review* 108 (3):735–62. doi: 10.1086/ahr/108.3.735.
- Tanner, Andrea, and Daniela B Friedman. 2011. "Health on the Web: An Examination of Health Content and Mobilising Information on Local Television Websites." *Informatics for Health and Social Care* 36 (1):50–61.
- Tewksbury, David. 2005. "The Seeds of Audience Fragmentation: Specialization in the Use of Online News Sites." *Journal of Broadcasting & Electronic Media* 49 (3):332–48. doi: 10.1207/s15506878jobem4903\_5.
- Thurman, Neil, and Steve Schifferes. 2012. "The Future of Personalization at News Websites: Lessons from a Longitudinal Study." *Journalism Studies* 13 (5-6):775–90.
- Weber, Matthew S. 2012. "Newspapers and the Long-Term Implications of Hyperlinking." *Journal of Computer-Mediated Communication* 17 (2):187–201. doi: 10.1111/j.1083-6101.2011.01563.x.
- Weber, Matthew S. 2014. "Observing the Web by Understanding the Past: Archival Internet Research." In *WWW'14 Companion Proceedings*. doi: 10.1145/2567948.2579213.
- Weber, Matthew S., and Peter Monge. 2011. "The Flow of Digital News in a Network of Sources, Authorities, and Hubs." *Journal of Communication* 61 (6):1062–81. doi: 10.1111/j.1460-2466.2011.01596.x.
- Weber, Matthew S., and Peter Monge. 2014. "Industries in Turmoil: Driving Transformation during Periods of Disruption." *Communication Research* 44:147–76. doi: 10.1177/0093650213514601.
- Westlund, Oscar. 2014. "The Production and Consumption of News in an Age of Mobile Media." *The Routledge Companion to Mobile Media*, 135–45. New York, NY: Taylor & Francis.

- Widholm, Andreas. 2016. "Tracing Online News in Motion: Time and Duration in the Study of Liquid Journalism." *Digital Journalism* 4 (1):24–40.
- Zimmer, Michael. 2015. "The Twitter Archive at the Library of Congress: Challenges for Information Practice and Information Policy." 2015. doi: 10.5210/fm.v20i7.5619.

## Appendix A: Online Source Search Protocol

The following outlines the key sources that were consulting in searching for news outlets within a given community.

### *Television*

Association of Public Television Stations' Station Directory  
FCC Broadcast Television License Database  
NPR Labs Mapping and Population System

### *Radio*

FCC AM and FM Broadcast License Database  
Radio Locator

### *Newspapers*

Library of Congress Directory of Newspapers – We decided that is something is listed as defunct by the Library of Congress' directory, then we would retain that notation.

### *Online News Sites*

Knight Foundation's Directory of Community News Sites  
Columbia Journalism Review's Guide to Online News Startups  
Online Newspaper Directory for the World  
Michelle's List

### *Alternative Sources*

Mondotimes

## Appendix B: Manual Source Search Protocol

In order to identify additional sources missed in the online search, a manual search protocol was developed in order to search for additional sources that were not readily accessible via traditional search.

1. Start with Wikipedia and search for each community
  - May list community's media
  - May provide community nickname(s) to utilize in search engine queries
  - May describe large minority populations that could be useful in search for minority/foreign language media outlets
2. Go to Patch.com
  - Enter community name into Find Your Patch pulldown menu
3. Google Search
  - Key terms:
    - "[Community Name][Community Nickname] News"
    - "[Community Name][Community Nickname] Journalism"
    - "[Community Name][Community Nickname] Hyperlocal"
    - "[Community Name][Community Nickname] Blog"

- Repeat in Spanish
  - News = Noticias
  - Journalism = Periodismo



# SAVING DATA JOURNALISM

## New strategies for archiving interactive, born-digital news

**Meredith Broussard** and **Katherine Boss**

*Important works of data journalism are disappearing from the web because they are too technologically complex to be captured or archived by libraries or web archiving technologies. Research based on journalism depends on the existence of news archives. For the benefit of future scholars, it is imperative that libraries and newsrooms solve this problem. This research contends that dynamic web archiving of data journalism will require a new, emulation-based approach to capturing these works. This new approach in turn necessitates new web archiving tools and workflows to enable collaborative collection of the projects, because unlike in print-based archiving, the process will depend on detailed technical information sharing among stakeholders.*

*Toward this end, this article summarizes the results of a questionnaire that described the most common frameworks, database technologies, and programming languages used to build 76 complex works of data journalism published between 2008 and 2017, as well the ways these works are being maintained and stored. This information can inform the development of emulation-based archiving tools to capture and preserve these stories using methods that would fit within the workflow of news organizations. This research is a first step toward devising an automated solution for long-term preservation of data journalism projects.*

### **Introduction**

Journalism research depends on the existence and maintenance of newspaper and magazine archives. Archiving the news has always been a difficult task, but the challenges to archiving online news are vastly different than those that have come before for several reasons. Digital news stories do not have the same clearly delineated boundaries as their print counterparts: stories in a physical newspaper have a clear start and end, but digital news stories are intertwined with the rest of the Internet in complex ways, such as when they contain images and videos that are served from other websites or that pull from an external API such as Google Maps. Defining the boundaries of a news website for a web archiving crawl can be quite complex. Similarly, the breadth of content that libraries need to archive is no longer straightforward. Libraries

used to save each edition of the published newspaper and all of its regional counterparts and sections. Today news organizations produce and distribute content in many different formats (print, online, broadcast) and on a variety of platforms (websites, social media, mobile applications, etc.), some of it redundant but much of it unique. While web archiving can capture a good deal of born-digital news, the dynamic nature of the web and the constantly changing platforms for distribution is making this an increasingly difficult task (Brügger 2011, 2018). Many of these platforms resist archiving or are guarded by terms of service that prohibit archiving. For example, libraries cannot easily archive the Facebook page of *The New York Times*, or the Instagram account for the *Washington Post*, because of these platforms' terms of service. Finally, online news has brought about issues with versioning. The "version" of each story to be archived used to be straightforward, in that the published print edition was the record of the news that day, and it was archived and preserved, with all of its typos and flaws, for history to judge. Scholars could trust that this news story was the one the public had seen and read. This has also changed; stories are now edited throughout the day, with few standards regarding how those edits must be acknowledged or recorded as they are made. News is also customized and A/B tested for different regions and demographics, and there is little way of knowing how many versions of a news story there have been, or which version most people would have seen.

For libraries, the challenges of archiving the total print and digital output of a news organization and of capturing a story throughout all of its iterations are formidable ones. Since the shift to digital, the archiving landscape has drastically changed, and so have the implications for research based on journalism. Scholars and historians often engage in content analysis of how a publication covers a subject over time, but researchers cannot currently access through online databases much of the digital content that a news organization has published in a year. Significant portions of the content are missing. This article explores the problems in archiving a specific form of journalism — data journalism — and raises questions about the infrastructure of archiving for the future of scholarly research. We ask: how has the widespread adoption of specialized data journalism practices created holes in the historical record that constitutes the foundation of journalism research, and what collaborative digital archiving strategies can be used to ameliorate the problem?

### Theoretical and Infrastructural Considerations

Scholars like Bowker and Star (2000) have shown that critical reflection on infrastructure, the "scaffolding in the conduct of modern life" (47), can yield insights into the invisible forces of classification and power that mediate relationships between materials and practices in sociotechnical systems. Following Ananny (2018), we conceive of digital journalism as an artifact created by a *networked press*, meaning "a digitally sociomaterial set of human and non-human actors held together in contingent, mutually shaping relationships" (115). Furthermore, we proceed from Braun's (2015) notion that studying the infrastructure of news distribution as a sociotechnical phenomenon can elucidate online news. Finally, we posit that data journalism, classified as a specialized sub-field situated within the larger field of journalism, is a site of great interest and importance to current and future scholars (Usher 2016; Anderson 2018).

Scholarly research about digital journalism depends on having a collection of digital news to study. In the past, scholars have defined the news website as a site of study (Brügger 2009), which is predicated on the notion that a website may be created that contains a complete archive of journalism content produced by a news organization at a specific point in time. When a communication researcher goes to a media organization's API or accesses a media organization's archive through a database in a library, the researcher expects that the content will be complete and will thus represent the complete corpus of work for the time period specified. If the corpus is incomplete, it damages the foundational assumptions of communication research. Scholars must be able to trust that archives are complete, and thus as scholars we begin by examining the processes that contribute to creating digital news archives.

### **Data Journalism and the Cultural Record**

Data journalism practices are diverse, and range from visualizations to database-backed stories (Usher 2016; Anderson 2018). Any potential methods must take one part of the problem at a time, so we begin our focus in archiving a specific type of data journalism story, often called an interactive news application. A news application is a type of data journalism story that is interactive, exploratory, and is often custom-built by a team of reporters and developers at a news agency (Boss and Broussard 2017). Examples of news apps include "The Color of Debt" by ProPublica, and "Gun Deaths in Your District" by *The Guardian*. Stories like these incorporate a database and allow readers to filter, search, and explore the data, to ask their own questions and find unique meaning in a story.

There is currently no canonical estimate of the number of organizations producing news applications, or the total number of news applications that have been published. However, some international and national surveys have been done in the United States, the United Kingdom, Norway, and Sweden that give an emerging picture of the volume and nature of this work being produced globally (Stavelin 2012; Appelgren and Nygren 2014; Howard 2014; Fink and Anderson 2015; Heravi 2017). These studies highlight the growing prevalence and importance of data journalism in newsrooms of all sizes and types.

### **Why Is Data Journalism Being Lost?**

Static web content built from text and image news stories is mostly captured by web archiving tools like the Internet Archive's Heritrix web crawler. But dynamic content, such as social media feeds or interactive visualizations, cannot be captured by these tools and as such is being lost (Hansen and Paul 2015; Boss and Broussard 2017; Hansen and Paul 2017). There are multiple reasons for this. The first is that the complexity of dynamic websites requires a new approach to archiving. This new approach will require technological tools, an archiving infrastructure, and a workflow between news organizations and libraries that doesn't currently exist. A short explanation of these issues follows.

Historically, the archiving technique libraries favor as the solution to the rapid format obsolescence of media innovations has been migration. Libraries migrate print

newspapers to microform, and then microform newspapers to digitized PDF/A files. They migrate 16 mm film reels to VHS tapes, DVDs, and streaming MP4 files. This strategy has limitations, but has been largely successful for physical materials and static digital objects like text, images, sound, or static websites (Von Suchodoletz and Van der Hoeven 2009). Web crawlers such as Heritrix were built with capturing the static Internet in mind. They are able to capture snapshots of the “front end” of a website, the graphically appealing part that is seen through a browser, but not the “back end,” the space on a web server where complex calculations or database interactions may take place. One of the only tools currently available to capture dynamic web content is *Webrecorder*, an open-source project by Rhizome (Kreymer and Espenschied n.d.). *Webrecorder* records networked traffic and processes within the browser while the user is interacting with the webpage, which allows it to capture complex Javascript and other dynamic elements. However, the archiving process can only be done manually. Every link of a website must be manually clicked on in order for archiving to occur. Thus, there is currently no scalable or systematic way to capture and archive these sites. The effort needed to simply maintain dynamic digital objects online, and the amount of time, work, and organization involved in migrating or upgrading the websites themselves and the software environment required for them to display and function is impractical, if not unfeasible. Software packages and computational platforms change vastly over time. For these reasons, digital archivists believe that to save dynamic content such as data journalism for the long term, we must emulate, not migrate, the object in its computational environment (Von Suchodoletz and Van der Hoeven 2009; Rechert et al. 2012; Johnston 2014; Rosenthal 2015).

To make the case for an emulation-based web archiving strategy, we can look at early data journalism projects and interactive news applications, which are also the main focus of this research. From an archiving perspective, news applications have multiple components that make them especially challenging to archive, including a database, the data in the database, the graphical interface that appears in the browser, accompanying text, and often images, videos, audio, and other multimedia components (Broussard 2015, 300; Klein 2012). If a news application is to be preserved, all of these components must be saved and reassembled in the same way, using the same software “stack” to ensure that the look, feel, and functionality of the story can be recreated. The reliance of any current piece of software on a series of software dependencies is often referred to as “dependency hell” for the frustration it causes. If one of the dependencies is missing or out of date, the project can break or fail to display.

If the site also requires browsers capable of interpreting specialized interactive content, such as Flash, this can also present huge problems. Most browsers available in 2018 do not have Flash Player software installed. When visiting sites built in Flash an error message appears instructing users to install the software to view the content. But Adobe, the maker and owner of Flash, announced that as of 2020, it will no longer update or distribute the Flash Player (Adobe Systems Incorporated 2017). This is a looming crisis for accessing older, interactive websites. Nearly, a decade of interactive news, including thousands of data journalism stories, were built in this now outdated software; *The New York Times* found in a content audit of their digital assets that a full 19 percent of their interactive projects, comprising more than 40,000 URLs, relied on Flash (Heideman 2017).

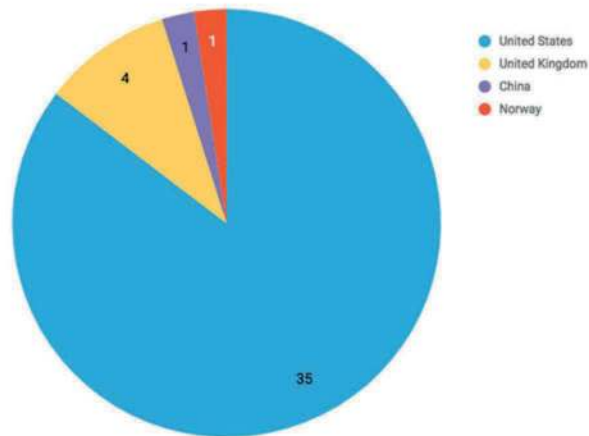
Preservation of floppy disks, CD-ROMs, video games and other virtual worlds, vintage software programs, and other dynamic systems that are more “closed,” comparatively speaking, than the internet, is currently being advanced by several institutions, including Rhizome, the Internet Archive, Carnegie Mellon, Yale, Deutsche Nationalbibliothek, and the British Library (Rosenthal 2015). These cutting-edge projects have not only captured and preserved the system images of these works, they have also advanced the emulator technologies and frameworks that allow users to find, access, and replay them on modern machines. However, none of this work has yet addressed emulation-based web archiving, which we propose here.

### Archiving via Emulation

Emulation-based archiving offers a viable solution to these problems by providing access to complex digital objects and interactive websites that migration cannot, and by addressing the rapid cycle of software obsolescence that migration cannot. In this process news organizations, and more specifically, system administrators or data journalists within news organizations, would be the first step in the archiving process. They would need to capture the system images of the data journalism projects including the files, programs, and data needed to recreate the website, as well as the software environment that enables it to function and display: the web browser, the operating system, etc. All of these files would then need to be packaged or compressed into an archivable and preservable format, or what archivists refer to as an “archival ingest package.” Once the files are packed they can live in dark archives or be sent, at scale, to institutions with the mandate, expertise, and funding to archive and preserve this content; namely, libraries and cultural memory institutions. Emulation is not a panacea, and it will require more economic resources, at least in the short term, than migration (Rosenthal 2015). But it is currently the most promising way to preserve dynamic journalism for the future (Boss and Broussard 2017).

This new approach to archiving via emulation will require a new workflow and collaboration between newsrooms and libraries or cultural memory institutions. This is because any emulation tool has to be deployed on a computer with access to the original software environment on which the website was built. Content creators must be willing and able to self-archive these works at the time of publication. The works would also need to be archived at multiple stages, as determined by curators and system administrators, so as to capture all of the relevant versions possible.

Given that there are hundreds of different programming languages, frameworks, and platforms that data journalists could utilize in their projects, descriptive data on the most common of these technologies will provide insights in designing effective preservation workflows and tools. This article summarizes the results of a survey that gathered data from national and international news organizations producing these stories, including *The Los Angeles Times*, *The Washington Post*, *The Guardian*, *The Wall Street Journal*, and ProPublica. Information was collected about the code, data, and server environments that make up these projects, as well as the proprietary and licensing information related to the data and editorial content. Our findings describe the set of most common languages and frameworks that make up 76 data journalism artifacts published in the past ten years, as well as the ways these works are being maintained



**FIGURE 1**  
Count of news organizations surveyed by country

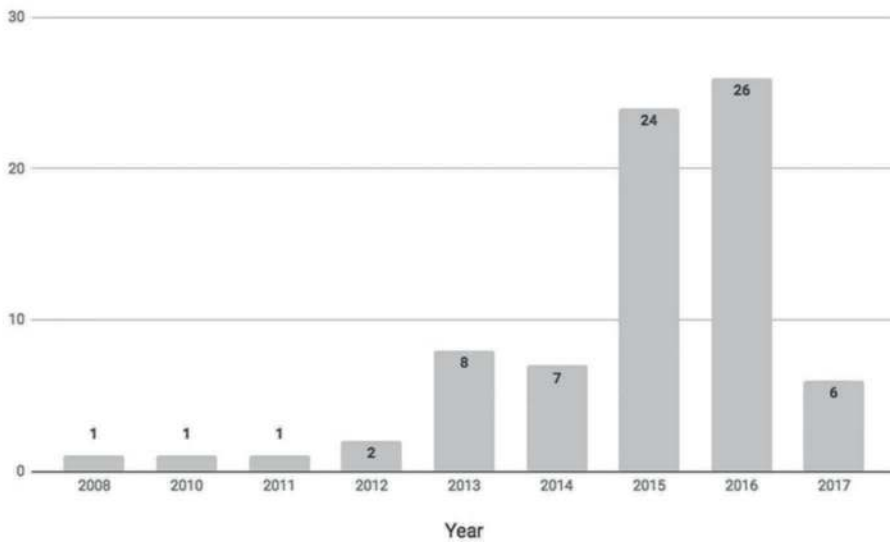
and stored. This information can be used to inform the development of tools to capture, archive, and preserve these vital data journalism stories so that they may be discoverable and accessible to future generations.

## Methodology

We created an online questionnaire in order to collect data on the code, data, software libraries, and server environment of news apps, as well as the proprietary and licensing information related to each app's data and editorial content. The purpose of this questionnaire was to take a snapshot of how news apps are being built and stored, to describe the licensing information of the editorial content and the underlying data, and to gather qualitative responses on the archiving process. We expect that the questions specific to news organizations and the code management and cloud hosting systems they use, the licensing information for their content, and their ability to archive dynamic content has elicited durable and replicable results. Since these projects continuously evolve, we do not expect that the specific technologies used to build individual news applications would be similar in a future survey of this kind; we include these data to document what is happening in this current moment.

The questions were based on the Performance Model Framework for the Preservation of a Software System (Matthews et al. 2010), which was previously identified as applicable to this research (Boss and Broussard 2017). The framework included the following metadata categories needed to properly describe, archive, and preserve a software object: functionality, software composition, provenance and ownership, user interaction, software environment, software architecture, and operating performance.

In order to ensure that data were collected from both small and large newsrooms, a series of structured interviews based on the questionnaire were conducted in the fall of 2016. For these interviews, we identified eight organizations of various sizes and types that are known for producing news apps. These were the following: National Public Radio, *The Texas Tribune*, *The Wall Street Journal*, the Center for Investigative Reporting, The



**FIGURE 2**

Year of first publication for each news application

Marshall Project, DataMade, Chalkbeat, *The Seattle Times*, and the Sunlight Foundation. We arranged a telephone interview with a staff member in each newsroom. A total of 21 structured interviews, each conducted by a graduate student, were collected.

Based on our analysis of the interviews, we adjusted some of the questions to elicit more detailed qualitative data. The revised survey consisted of 40 questions that gathered descriptive data about the way news apps are built and stored. We distributed this revised version in the fall of 2017 via e-mail, on social media, and via the list-serv of the National Institute for Computer-Assisted Reporting (NICAR), which is the largest US-based online mailing list of data journalists.

Several respondents completed the questionnaire multiple times in order to describe multiple news apps, and one incomplete response was discarded from the sample. Since some of the questions were specific to individual news applications, and others were specific to the news organization as a whole, the analysis of the responses reflects these different totals respectively.

## Results

A total of 76 complete responses from 41 media organizations were gathered. The majority of responses came from news organizations in the United States and the United Kingdom (see Figure 1). Although there is no canonical or definitive number of the total population of news organizations producing data journalism internationally, a recent global survey suggests that this is a relatively small community (Heravi 2017). We, therefore, estimate that this sample size is sufficient to represent durable and replicable results at the organizational level on the state of data journalism archiving in the United States.

More than 90% of the news apps in the study were first published within the past 5 years, with the oldest news app dating back to 2008 (see Figure 2).

**TABLE 1**  
Technologies used to build each news application

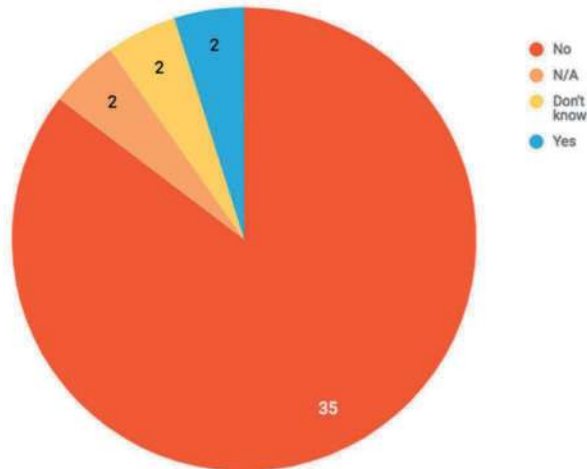
Programming language	Framework	JavaScript library	Other
Ruby	Django	jQuery	Sass
R	Bootstrap	D3	Grunt.js
Python	Backbone	Leaflet	Django-bakery
SQL	Flask	Underscore	Gulp
JavaScript	Rails	Ractive.js	Elasticsearch
HTML/CSS	Ruby on Rails	Raphael	Mapbox
	Node	React	Babel
	Angular	Landline.js	Bower
	WordPress	Tabletop	Frozen-Flask
	Express	Mapbox-gl.js	Google Maps API
	Tablestacker	Datatables	
		Highcharts	
		Tachyons	
		Velocity.js	
		Chartz	

The questionnaire revealed the set of frameworks, libraries, database technologies, and programming languages that are ubiquitous in the current data journalism realm. In response to the question, "What programming languages did you use to build the app?" respondents indicated that 53 used JavaScript, 25 used Python, 13 used a variety of SQL, nine used Ruby, and one used R. The varieties of SQL mentioned were PostgreSQL and SQLite. Mapping technologies represented in the sample included PostGIS and the Google Maps API. All of the data journalism projects identified in the survey used HTML/CSS because all were web-based. Table 1 contains a list of the most common technologies used.

All of the applications surveyed stored their code in a code management system. The code for more than half of the news apps surveyed was managed in Github (53 of 76); other common responses were Bitbucket and Unfuddle. These repositories are spaces where programmers store code as they develop it in order to manage collaborative issues such as bug tracking, task management, and version control. However, while the source code can exist within a code management system as long as the account is not deleted by the owners, this code will be useless when the environment designed to run it has become obsolete and cannot be recreated. It must be emphasized that merely maintaining a repository for the source code does not ensure that the website will be functional in the future. A code management repository does not constitute a digital archiving system.

Very few of the respondents reported having an archiving system for their interactive projects. In response to the question, "Does the news organization have an archiving system for news apps?" only two of 41 organizations, or 5%, answered "yes." In a follow-up conversation with a team of developers and editors at one of these organizations, we discovered an extremely fragile, ad hoc archiving strategy that lacked documentation, was maintained and understood by only one person, and did not address any aspect of digital preservation or impending software obsolescence.





**FIGURE 3**

Count of responses to the question, "Does the news organization have an archiving system for news apps?"

Fully 85% of the news apps described in this survey were not being archived anywhere (see Figure 3). Comments related to this question underscored how time is scarce in newsrooms, and archiving an afterthought:

"When news apps break because tech changes, they don't usually take the time to fix it unless it's really simple,"—*Non-profit US news organization*

"No archives!"—*Commercial US news organization*

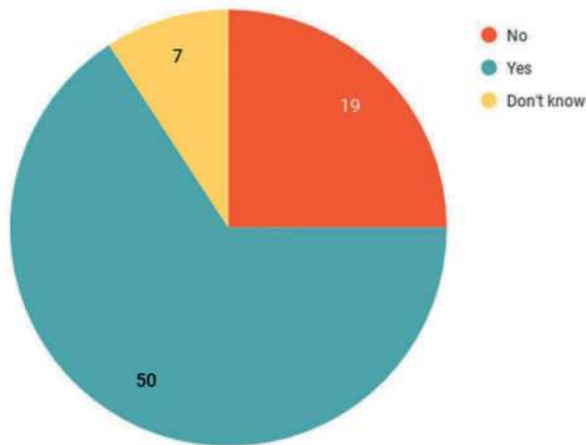
This alarming finding relates to the fact that there are few, if any, staff in modern newsrooms who wake up each day thinking about digital archives and preservation. This is a direct result of the shuttering of newsroom libraries, a trend that began in the early 2000s when budget cuts and the disruption of the publishing industry led newsrooms to close their libraries and lay off their archivists (Hansen and Paul 2002, 2017).

In lieu of long-term archives, regular maintenance and updates to a news app could prevent these sites from breaking or becoming inaccessible. In response to the question "Is the news app being regularly maintained or updated?" for 25% of news applications (19 of 76) the response was "No;" for another 9% of news applications (7 of 76) the response was "Don't know." Given that 93% of the apps were first published in the past 5 years, this is an indication that there is little institutional support for maintaining older digital work (see Figure 4). Comments related to this question underscored how much effort is required to maintain the project for the long term:

"Project is broken. I am fundraising to pay for tech help to fix it."—*Independent US data journalist*

Almost half of the applications were hosted on Amazon Web Services (49%); other common cloud hosting services used were Google Cloud, Digital Ocean, Heroku, or a private server managed by the newsroom (see Figure 5).

We also asked respondents to describe the copyright of the data used in each news app. Slightly more than two-thirds of the news apps described (approximately 68%) used publicly available data, most of it gathered via Freedom of Information Act



**FIGURE 4**

Count of responses to the question, "Is the news app being regularly maintained or updated?"

(FOIA) requests or scraped from the public APIs of government websites. The remaining third of news apps used data that was either proprietary (12%) or a combination of public and proprietary (9%); 12% of responses to this question were "Don't know."

### Industry Needs

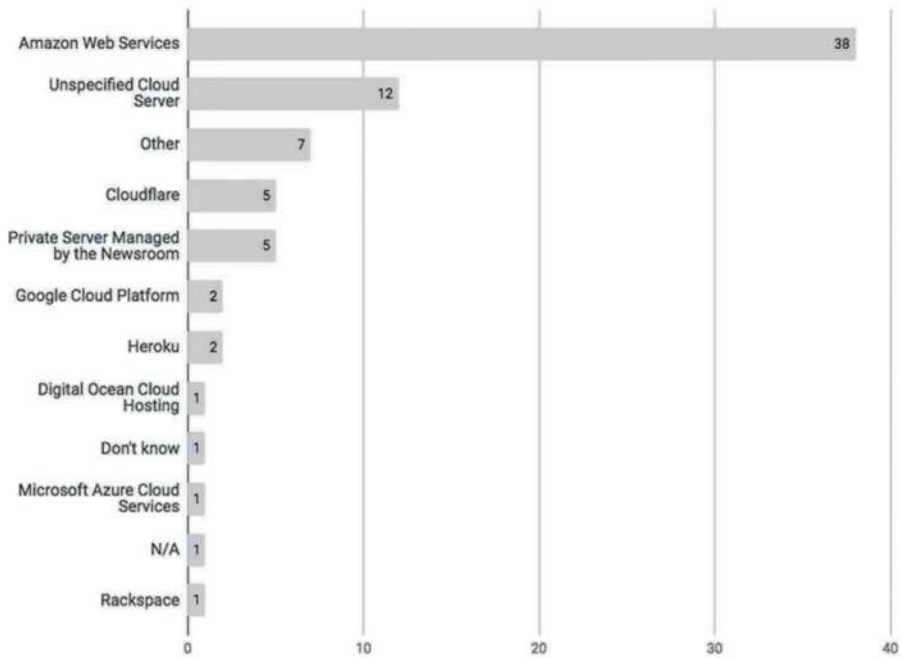
The findings from the questionnaire led us to draw conclusions about key areas of need for archiving, including insights about archiving processes that would fit within the organizational workflow of news organizations.

#### *Need for an Emulation-Based Web Archiving Tool*

Several tools are in development that could facilitate an emulation-based web archiving process and compress data journalism projects into a single archivable ingest package. One tool with a promising level of applicability to archiving news apps is the open source, computational reproducibility software *ReproZip*. Designed to make computational experiments reproducible across different platforms and over time, *ReproZip* could be customized for the specific needs of news apps (Chirigati et al. 2016). The authors are currently investigating this option through a research project funded by the Institute of Library and Museum Services, but more work is needed on using virtual machines and reproducibility tools to pack and archive data journalism projects through emulation (Institute of Museum and Library Services 2018).

#### *Need for Archiving Incentives and Workflows*

Data journalism is costly and time consuming to produce, and in resource-depleted newsrooms archiving is often severely underfunded, if it is funded at all. News archiving has always been seen as a drain on precious resources in newsrooms, and is rarely viewed as financially viable (Hansen and Paul 2017). But the loss of the historical news record is priceless, so stakeholders must look for ways to financially



**FIGURE 5**

Hosting provider used to serve each news application

incentivize organizations to begin the process of archiving their work. Currently, many newsrooms are able to leverage and monetize the archives of their traditional print content by licensing it to information vendors like ProQuest or LexisNexis. These vendors in turn sell and license that content to libraries, other newsrooms, and businesses. Through this process, the news is made archivable, preservable, and discoverable in a way that is also economically feasible for newsrooms. A similar monetization strategy could be pursued to leverage newsrooms' born-digital and interactive content.

In newsrooms, time is also a precious resource. Just as in the print world, data journalists tend to publish their stories and then move on to the next project. As this survey confirmed, data journalists don't tend to provide long-term support for their projects, and media organizations offer few institutional resources for maintaining older digital work. This reflects organizational priorities. Journalists optimize for the news cycle, which thrives on novelty and frequent releases. The greatest number of users will pay attention to the app at the time of the app's release, then user attention will decline over time. Since an emulation-based archiving tool must be deployed at the time of publication, this research suggests that the tool would need to be seamless, simple, and quick, taking up as little time and as few resources as possible. This would increase the chances that emulation-based archiving would be adopted widely in newsrooms where time and money are scarce.

### *Need for Improved Communication and Consistent Terminology*

In this research, we also observed a major disconnect between how librarians and data journalists talk and think about archiving. When librarians talk about archiving

a digital object, they refer to a process that enables the long-term storage, preservation, and access to the materials (Hodge 2000). This process includes capturing the fundamental elements of the object to ensure the “look and feel” can be maintained, providing a permanent location and storage site for the work, describing the work with technical, bibliographic, and preservation metadata, and allowing access through a discovery or retrieval system. In contrast, when the data journalists surveyed in this study mentioned an archive or archiving, it became clear that many considered backing up on Github to be digital archiving. It is not. The code saved in a code management system is useless if the digital environment in which it was built no longer exists. Saving the code is just one piece of the puzzle; much more information and context is needed to archive a website, much less preserve it for the long term. This is another obstacle to saving these works. Improved communication is needed between newsrooms and library and memory institutions to better discuss the problem and potential logistical and technological solutions.

Respondents to the survey also talked about data journalism artifacts in different ways, and their language was inconsistent, but taxonomy is crucial to long-term archiving. Identifying a digital journalism project as a specific type will allow archivists to collect all of the necessary information for preserving that type of digital object. A coherent, consistent description of a news app is needed, particularly one that would differentiate between different types of interactives: news apps, data visualizations, infographics, etc. To further this discussion, we have identified four unique components of a news application:

1. The front end, or user-facing component. This is generally accessed through a web browser. It may consist of static or dynamic HTML/CSS.
2. The back end, or server-side component. In a database-driven news app, the database is considered part of the back end. Data processing may also take place on the server side depending on the choices made by the developers of the news app.
3. Structured data. The front end and/or the back end interact with the data. The data may be in a static files or a database hosted by the news org; alternatively, it may be hosted by a different platform and accessed via API.
4. The server environment. The web is a client-server environment. The client, a web browser, interacts with code that sits on a server. Code can't run on just any server; it is optimized for a specific hardware and software configuration. App-specific server configuration allows the app to read the structured dataset.

### *Need for Conversations Around Copyright and Intellectual Property Issues*

Code, like writing, is considered original work and is thus protected by copyright. However, many technological innovations are built on code developed by others. This has led to a complicated intellectual property environment. Some organizations apply the MIT License to their code, giving unlimited usage rights to all who are interested as long as the license travels with the code. Other organizations provide the code for reuse with the stipulation that the product not be used to promote anything without

their express permission. Our findings indicate that for the majority of news apps the code resides in private repositories, accessible only to members of an organization. Most news apps are proprietary, meaning that the software belongs exclusively to a person or organization and is not shared. This may have to do with organizational priorities, or it may have to do with the fact that preparing code for open release is extremely time-consuming. At the end of a large, computationally intensive project, data journalists are often so exhausted by launching the project that they do not prioritize the additional work required to clean the code and bring it up to the media organization's high editorial standards.

The editorial content of each news app is generally owned or copyrighted by the publishing media organization. This presents a challenge to archivists in that an archiving solution will require consent from the news organization. Code and editorial copyright will have to be negotiated so that memory institutions can preserve the full content and code environment of a news app. A best case scenario for the United States would be a legislative mandate, similar to those in Scandinavian countries, for legal deposit of born-digital works (Schostag and Fønss-Jørgensen 2012).

### Discussion and Future Research

For the past 250 years, the United States has struggled with the question of who holds the responsibility for archiving the news, and has faced a number of major challenges in trying to preserve its journalism in all formats. The currently available archives for print, visual, and broadcast news are just a fraction of what has actually been produced and published; a staggering amount of journalism has been lost. There are many different reasons for these losses. In Hansen and Paul's comprehensive history of news archiving, *Future-Proofing the News* (2017), some major themes emerge: news organizations that lack time and money for archiving and have little financial incentive to do so, an absence or delay in a legal imperative to archive content via copyright or legal deposit laws, libraries that cannot get access to news content to begin the archiving process or situations in which the law actually thwarts libraries from doing so, and limited library funding for archiving and long-term preservation. And finally, the issue that all variants of the news have faced over time: each new media format precipitates a process and the technology by which to archive and preserve it. These processes are part of infrastructural considerations that constitute the complex interplay between materials and practices in sociotechnical systems. Human considerations such as legal and copyright issues and non-human actors such as hardware and software constraints are interdependent, change frequently, and are crucial to understanding the complex environment of digital news.

Now as ever, newspapers are undergoing a rapid and continuous evolution of journalism production, publishing, and distribution. This digital shift is a rich site of study for current and future communication scholars. However, scholarly work proceeds from the assumption that the scholar is able to access a robust corpus of artifacts. It is not currently safe to make this assumption, given that newsrooms and libraries cannot currently archive dynamic digital news content at scale in any way. The results of this questionnaire describe 76 news applications published by 41 different organizations since 2008, and identify some of the dominate technologies, hosting providers, code management

systems, and intellectual property licenses for these works. Of the organizations surveyed, only two had any sort of archiving system for their news applications, neither of which are taking measures to ensure long-term access to the works.

This is a crisis that affects the future of journalism-based research. Newsrooms are volatile organizations that can be quickly bought and sold, founded, and shuttered. During the year-long course of this research, one of the respondents to the survey, a non-profit organization aimed at improving government transparency, ceased its data journalism operations and turned over control of its projects to other organizations. The URL of the news application described was already inaccessible as of this publication. It stands as a painful reminder of the urgency of this problem.

The two important stakeholders needed to establish legal agreements and workflows to successfully begin saving the content remain the same: news organizations and libraries or cultural memory institutions. Cooperation and coordination between these groups has always been the way to archive news content at scale, and in the case of data journalism and interactive news content this remains true. The responsibility for archiving must be felt by both groups.

The data gathered from this questionnaire confirms the extent of this problem and points to a need for a significant shift in web archiving strategies from static to dynamic archiving. It is imperative that newsrooms, libraries, and cultural memory institutions work together to find a solution for capturing and archiving these works as soon as possible. First and foremost, this will require an emulation-based web archiving tool. This software or another future software intervention will ensure that news organizations can save their work so that we can read today's news on tomorrow's computers.

## ACKNOWLEDGEMENTS

The authors are grateful to Eva Revear for her contributions to this project, and the organizers and participants of the Dodging the Memory Hole conference who have provided valuable feedback on this research.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

## REFERENCES

- Adobe Systems Incorporated. 2017. "Flash & The Future of Interactive Content." Adobe Blog (*Blog*). July 25, 2017. <https://theblog.adobe.com/adobe-flash-update/>.
- Ananny, Mike. 2018. *Networked Press Freedom: Creating Infrastructures for a Public Right to Hear*. Cambridge, MA: The MIT Press.
- Appelgren, Ester, and Gunnar Nygren. 2014. "Data Journalism in Sweden: Introducing New Methods and Genres of Journalism into 'Old' Organizations." *Digital Journalism* 2 (3): 394–405. <https://doi.org/10.1080/21670811.2014.884344>.
- Anderson, C. W. 2018. *Apostles of Certainty: Data Journalism and the Politics of Doubt*. New York, NY: Oxford University Press.

- Boss, Katherine, and Meredith Broussard. 2017. "Challenges of Archiving and Preserving Born-Digital News Applications." *IFLA Journal*, 43, 034003521668635. <https://doi.org/10.1177/0340035216686355>.
- Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences. Revised edition*. England: The MIT Press.
- Braun, Joshua A. 2015. *This Program Is Brought to You By ...: Distributing Television News Online*. New Haven: Yale University Press.
- Broussard, Meredith. 2015. "Preserving News Apps Present Huge Challenges." *Newspaper Research Journal* 36 (3): 299–313. <https://doi.org/10.1177/0739532915600742>.
- Brügger, Niels. 2009. "Website History and the Website as an Object of Study." *New Media & Society* 11 (1–2): 115–132. <https://doi.org/10.1177/1461444808099574>.
- Brügger, Niels. 2011. "Web Archiving – Between Past, Present, and Future." In *Handbooks in Communication and Media: The Handbook of Internet Studies*, edited by Mia Consalvo, and Charles Ess. New York: Wiley. [https://ezproxy.library.nyu.edu/login?url=https://search.credoreference.com/content/entry/wileyhins/web\\_archiving\\_between\\_past\\_present\\_and\\_future/0?institutionId=577](https://ezproxy.library.nyu.edu/login?url=https://search.credoreference.com/content/entry/wileyhins/web_archiving_between_past_present_and_future/0?institutionId=577).
- Brügger, Niels. 2018. "Web History and Social Media." In *The SAGE Handbook of Social Media*, 196–212. London: SAGE Publications Ltd. <http://sk.sagepub.com.proxy.library.nyu.edu/reference/the-sage-handbook-of-social-media/i1587.xml>.
- Chirigati, Fernando, Remi Rampin, Dennis Shasha, and Juliana Freire. 2016. *ReproZip: Computational Reproducibility With Ease, 2085–2088*. San Francisco, USA. <http://big-data.poly.edu/~fchirigati/papers/reprozip-sigmod206.pdf>.
- Fink, Katherine, and C. W. Anderson. 2015. "Data Journalism in the United States: Beyond the 'Usual Suspects.'" *Journalism Studies* 16 (4): 467–481. <https://doi.org/10.1080/1461670X.2014.939852>.
- Hansen, K. A., and Paul, N. 2002. "Reclaiming News Libraries." *Library Journal* 127 (6): 44.
- Hansen, K. A., and N. Paul. 2015. "Newspaper Archives Reveal Major Gaps in Digital Age." *Newspaper Research Journal* 36 (3): 290–298. <https://doi.org/10.1177/0739532915600745>.
- Hansen, Kathleen A., and Nora Paul. 2017. *Future-Proofing the News: Preserving the First Draft of History*. Lanham: Rowman & Littlefield Publishers.
- Heideman, Justin. 2017. "URLs Should Never Die; Retiring Old Technology While Preserving The New York Times' First Draft of History." Presented at the Dodging the Memory Hole 2017: Saving Online News, San Francisco, CA, November 16.
- Heravi, Bahareh R. 2017. "The State of Data Journalism Globally." Proceedings of the First European Data & Computational Journalism Conference, Dublin, Ireland, 6 July 2017.
- Hodge, Gail M. 2000. "Best Practices for Digital Archiving: An Information Life Cycle Approach." *D-Lib Magazine* 6 (1). <https://doi.org/10.1045/january2000-hodge>.
- Howard, Alexander. 2014. "The Art and Science of Data-Driven Journalism." Tow Center for Digital Journalism.
- Institute of Museum and Library Services. 2018. "Grants Awarded: LG-87-18-0062-18." April 18, 2018. <https://www.imls.gov/grants/awarded/lg-87-18-0062-18>.
- Johnston, Leslie. 2014. "Preserving News Apps | The Signal." Webpage. March 11, 2014. [blogs.loc.gov/thesignal/2014/03/preserving-news-apps/](http://blogs.loc.gov/thesignal/2014/03/preserving-news-apps/).
- Klein, Scott. 2012. "News Apps at ProPublica." In *The Data Journalism Handbook*, edited by Jonathan Gray, Liliana Bounegru, and Lucy Chambers, 1st ed. Sebastopol, CA: O'Reilly Media. [http://datajournalismhandbook.org/1.0/en/delivering\\_data\\_2.html](http://datajournalismhandbook.org/1.0/en/delivering_data_2.html).

- Kreymer, Ilya, and Dragan Espenschied. "Webrecorder: A Project by Rhizome." Accessed January 10, 2018. <https://webrecorder.io/>.
- Matthews, Brian, Arif Shaon, Juan Bicarregui, and Catherine Jones. 2010. "A Framework for Software Preservation." *International Journal of Digital Curation* 5 (1): 91–105. <https://doi.org/10.2218/ijdc.v5i1.145>.
- Microform. 2014. Encyclopaedia Britannica.
- Rechert, Klaus, Isgandar Valizada, Suchodoletz Dirk von, and Johann Latocha. 2012. "BwFLA – A Functional Approach to Digital Preservation." *PIK – Praxis Der Informationsverarbeitung Und Kommunikation* 35 (4): 259–267. <https://doi.org/10.1515/pik-2012-0044>.
- Rosenthal, David S. H. 2015. "Emulation & Virtualization as Preservation Strategies." [https://mellon.org/media/filer\\_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf](https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf).
- Schostag, Sabine, and Eva Fønss-Jørgensen. 2012. "Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective." *Microform & Digitization Review* 41 (3–4): 110–120. <https://doi.org/10.1515/mir-2012-0018>.
- Stavelin, Eirik. 2012. "Nyhetsapplikasjoner: Journalistikk Møter Programmering." In *Nytt På Nett Og Brett: Journalistikk i Forandring*, edited by Martin Eide, Leif Ove Larsen, and Helle Sjøvaag, 107–125. Oslo: Universitetsforlaget.
- Usher, Nikki. 2016. *Interactive Journalism: Hackers, Data, and Code*. Urbana: University of Illinois Press.
- Von Suchodoletz, Dirk, and Jeffrey van der Hoeven. 2009. "Emulation: From Digital Artefact to Remotely Rendered Environments." *International Journal of Digital Curation* 4 (3): 146–155. <https://doi.org/10.2218/ijdc.v4i3.118>.



# THE POLITICS OF WOMEN'S DIGITAL ARCHIVES AND ITS SIGNIFICANCE FOR THE HISTORY OF JOURNALISM

**Pernilla Severson**

*This article explores the politics of digital archives focused explicitly on women journalists and their work. A key question is here the wider implications and value for journalism historiography. A qualitative analysis is conducted of the online presence of two illustrative archives, one an oral history project called Women in Journalism and the other a women's history database called Kvinnsam. The analysis finds that whereas the archives do not lend themselves to participation as agency in co-constructing history, they give access to otherwise nonsearchable, nonvisible, and nonaccessible material of relevance to the history of women journalists and their work. The agency and political power of the archives are dependent on institutions, first, to simply materialize as online archives and, second, to (potentially) affect political matters and express political acts of resistance. For journalism history studies, this means engaging with the archives that exist, what forms they have, and how they are used. For digital journalism, this also implies a discussion of how archival experimenting could develop the field.*

## Introduction

A recurrent issue for journalism history relates to the relatively narrow range of sources used by journalism historians: news media texts or the personal papers of individual journalists (Nerone 2011). A related issue is the call for journalism history studies to more actively write women into history (Steiner 2017). One problem is that certain types of sources lend themselves to telling particular types of stories (Nerone 2011); another is that a journalism history that excludes women's perspectives creates a distorted perception of what journalism is and has been (Beasley 2010).

Following this, archives should be seen as political acts for advancing the history of women's journalism (Tusan 2005). Feminist historiography as the study and creation of feminist archives to promote gender equality is growing (Cifor and Wood 2017; Eichhorn 2014). It also has a history. In 1935, the World Center for Women's Archives was initiated by building a collection with the intent of its being a counter-archive using alternative approaches to represent women's experiences more broadly. This shaped feminist historiography, the women's archive movement and archival

scholarship, particularly regarding marginalized groups (Lubelski 2014). This is linked to proactive collecting, such as oral history projects providing inadequately documented groups a voice and the formation of women's archives to enable collection development policies (Zanish-Belcher and Mason 2007).

Digital archives could broaden their range of sources by including women's perspectives but could also simply continue and thus increase bias. The politics of archives and digital archives are part of a broader research context giving voice to or silencing marginalized communities. The archive is recognized as an incomplete repository, with silences, gaps, and elisions (Thomas, Fowler, and Johnson 2017). The politics of archives is researched in various fields (Casswell 2014; Derrida 1996; Hoskins 2017; Robertson 2011). Postcolonial research shows, for example, how the archive is a co-creator of the forgotten history of oppressed peoples (Burton 2003; Gauthereau 2017). The focus in this study is on the politics of women's digital archives in relation to journalism history studies.

The digital archive suggests "a new kind of archive, with new structures, new ways of searching, a paradigm shift in record keeping" (Johnson 2017, 154). Marginalized communities can be co-creators with archivists in selecting material and designing interfaces as seen in community archives (Johnson 2017). Eichhorn (2014) places archives in a feminist and activist movement as way of enhancing agency and power where digitization matters, including roles for feminist action as a radical cataloger or as an accidental archivist. In contrast, there is a fear that digitization only happens for fields and topics that are already popular. Uricchio (2014) makes a case for contours of absence in the construction of media history and the need to "make productive use of the historiographic problems we face" (126).

Archive politics suggest that one of the roles of the archive and the digital archive is to give voice to women journalists and to engage in a critical understanding of how to think about digital media, history and gender. Therefore, the purpose of the study is to analyze and distinguish what voices are made present in two illustrative women's digital archives for journalism history, how digitization matters in this voice-making, and how this can be understood in relation to democratic values.

## Archives

While digitization has made scholars from various disciplines interested in how archives shape an interdisciplinary field there is still a lack of agreement on what counts as an archive/digital archive. Archival researchers define archives as records created by a social actor (individual, institution, or organization) in a process through which they are preserved due to their permanent value (Theimer 2012). The unit of the archive is the document: any discreet piece of information (Howell 2006). Digital humanists understand archives as selections, consisting of clustered online material, which can comprise both digital and digital copies of analog material. The selection often consists of materials located elsewhere, such as physical repositories or collections. The archive then means a selection of purposefully collected material. For an organization, an archive is often the place to retain and organize records of the organization (Theimer 2012), like the online news archive of a daily newspaper.

An archival field relevant for the understanding of archives and digital archives is what Theimer (2012) calls participatory archives, which are new forms of archival activity: “an organization, site or collection in which people other than archives professionals contribute knowledge or resources, resulting in increased appreciation and understanding of archival materials and archives, usually in an online environment”. Digitization can provide online access to previously analog material as well as other forms of field collection with participatory approaches (Theimer 2012).

A part of archival knowledge building is digital historiography: a research approach for studying the interplay between digital technology and historical practices focusing on contextual implications (Theimer 2012). Digital archives are understood as everything from traditional physical archival materials represented digitally to born digital materials. Digitally represented traditional physical materials include descriptions in online finding aids and catalog records. Collections of digitized analog historical materials can also be seen as forms of digital archives as repositories that may give online access to digitized collections. A born digital archive is, for instance, the selected digital files from Salman Rushdie’s Macintosh Performa 5400, as well as The September 11 Digital Archive with a crowdsourced collection of materials related to the attacks of 9/11. Born digital archives are also archival initiatives, such as “Web Archives,” which harvest content from the web for long-term access and preservation (Theimer 2012). These archives function as sources for history-writing with Web-based materials (Rogers 2017).

### **The Selected Archives**

Complete inventories of digital archives or of women’s journalism archives, analog or digital, are nonexistent. There are initiatives online, like ArchiveGrid (2018) or Wikipedia (2018), and random selections of women’s history archives, like Centre d’Archives et de Recherches pour l’Histoire des Femmes (2018). However, these initiatives are incomprehensive and biased toward the Western world. This article is based on an exploratory qualitative study aimed to initiate and accumulate knowledge of women’s digital archives for journalism history studies. Thus, a representative selection is neither possible nor desirable.

The studied archives have been chosen for the different ways they raise (theoretical) questions about archives and the history of women journalism/journalists. The selected archives are compared and contrasted in order to produce a deeper and richer picture of each archive.

Women’s journalism history research shows how daily news press and journalism unions have been powerful archive initiators and builders. However, for women’s journalism history the newspapers’ own archives are incomplete and lack relevant tagging. Important sources have, therefore, been micro-filmed newspapers at libraries as well as material from women’s research databases (Ney 2006). Material from these databases is primarily historical books on early female journalists, some in the U.S., and then often the result of academically educated female journalists’ research (Ney 2006).

I also strategically probed online information to find relevant digital archives and selections. I explored digital archives by searching for “women,” “journalist,” and “archives” in various Asian and African languages using google translate and the

google search function through all results. No archives were found. I found media history archives digitizing analog material, like *The Interviews: An Oral History of Television!* (2018). It is quite common in the U.S. to use oral history to capture the legends of particularly important women. Another example of women in journalism archives is *The Herstory: JAWS Oral History Project* (2018). An example from popular culture is *The Women Who Rock Digital Oral History Archive* (2018), where the digital includes co-creating the archive in various ways. The 1947 Partition Archive (2018) on the partition of British India to India and Pakistan is another example of more traditional cultural heritage approaches: crowd-sourcing of partition witness interviews where volunteers are trained in the oral history technique. Within journalism and women journalism, this is nonexistent.

A particular women's journalism oral history project is *Women in Journalism* (2018) (WiJ). Women's National Press Club, an organization founded to support equal rights for women in the newsroom, initiated the project in 1986. Since 1987, "full-life interviews" (life history interviews) have been collected. WiJ consists of 68 interviews of women journalism pioneers. The self-description of the project states that:

The collection is an important part of the history of journalism as well as showing a very interesting perspective on the history of women in the workplace. As the collection is digitized it will be available for use by scholars to further their research and to educators for development of courses on journalism history and women's studies. (WiJ)

WiJ began the project "Archive Digitization" by digitizing the Cora Rigby Archives and the *Women in Journalism Oral History* project materials. WiJ has been awarded for its achievements in presenting women's journalism history and has been the subject of several studies (Beasley 2015; Fuchs 2003; Whitt 2008). I selected WiJ for this study due to all these traits.

After examining various forms of digital archives, I decided to select a dominant digital archive initiative of women history, a women's history archive. I would preferably have wanted to include born digital archives or oral history-based archives in two contexts (countries). I have, however, not been able to locate digital women's journalist archives with aspects resembling community archives or "more digital" archives. There are, of course, digital archives preserving the digital without an explicitly focus on women, e.g. *The Journalism Digital News Archive* (2018), an online archive of news content in digital format. In my country, Sweden, there are no oral history archives for women journalists as well as for other media professions or journalist-related aspects and themes.

Based on these considerations I selected *Kvinnsam* (2018), an archive that is a repository and database with accessible digital documentation of analog material, including particular collections. *Kvinnsam* is the database component of an established library search system made in cooperation with the Secretariat for Gender Research, at Gothenburg University in Sweden. *Kvinnsam's* began with the *Women's History Collections*, founded as a private initiative in 1958. Since the mid-1980s, the collections have had their own premises. *Kvinnsam's* cataloging of new literature is based on the collections of the university library. The collections consist of books, journals, articles, chapters, pamphlets, research reports, etc. *Kvinnsam* has been online since August

1998 and is available via LIBRIS (Library Information System of Sweden). The database is in Swedish and English. Kvinnsam is, in turn, also part of a larger collaboration with Nordic and European women's and gender archives. Kvinnsam is a well-known and legitimate actor that illustrates how women's archives can be created and developed.

The similarities and differences between the two chosen archives arguably create a favorable starting point for an analysis with the ultimate aim of contributing to increasing the presence, power and value of digital archives for women's journalism history studies.

### **A Qualitative Archive Analysis**

This is a qualitative archive analysis where the two archives have been chosen for the different ways they raise (theoretical) questions about archives and the history of women journalism/journalists. More specifically, theoretical propositions help generalize from the archives as analytical generalizations. The analytical technique is pattern matching between the archives as well in relation to theoretical propositions that take into consideration rival patterns.

The theoretical propositions chosen for this study are affordance theory and voice as participation and power. The affordance perspective in this study focuses on the relational aspects of the social and the material; and this perspective will be combined and furthered using other theories (voice, participation, and power). This means that affordance theory directs attention to the potential digital characteristics (meaning the affordances) of the digital archives while the theories of voice and participation as power function as "lenses" which allow a discussion of the complicated problems and social issues of the politics of women's digital archives focused explicitly on women journalists and their work. In this context affordance theory is a sensitizing concept for focusing on and identifying what a digital archive could be, and then voice, participation, and power are theories to discuss archival politics.

An affordance approach is a way to meaningfully structure an analysis of the relationship between technology and the social. Affordance theory is a micro-level theory on the very specific relationship between the social and the technical. Gibson (1979) developed affordance theory to explain the relationship and complementarity of an animal and its environment, naming affordances as a form of action possibility. To adopt an affordance perspective is to recognize both use and how an object's materiality could invite and constrain this use. Hence, affordance theory is not only to be understood as interface and technical affordances connected to a device's interface. The affordances are located both in the social and in the technical. Depending on interest you can emphasize either the social or the technical. Design research and Human-Computer-Interaction research emphasize the technical design of affordances. For a review of how the term affordances has been used by communication scholars, see Nagy and Neff (2015).

In this study, affordance theory guides the selection of relevant study object (digital archives) and help to interpret possible forms of usage. Similar work in journalism research has, for example, been done by Tenenboim-Wenblatt and Neiger (2018) in their development of temporal affordances to study the relationship between news as technology and journalistic storytelling practices. Another example is Djerf-Pierre,

Ghersetti, and Hedman (2016) use of affordances to avoid static conceptions of both uses and technologies in studying journalists' appropriation of social media affordances.

Simply using identified affordances in an analysis, creates an emphasis on discovery and description. This problem with the affordance approach has recently been discussed by media scholars. Shaw (2017) shows the need to link affordance theory to other theories and merges it with Hall's dominant/hegemonic, negotiated, and oppositional reading positions to approach the political implications of audience activities with these technologies in new and more nuanced ways. In this study affordances aid the understanding of an archive's repertoires of action (Basu and de Jong 2016). Affordance as theory provides insight into potential digital archive characteristics to gain an understanding of archival properties that relate to specific usages as well as a range of possible developments.

I mainly build my work on affordances on Evans et al (2017) that show how affordances need to meet three criteria: that an affordance is neither the object nor a feature of the object (features are static while affordances are dynamic, a table is the object and eating is the affordance); that the proposed affordance is not an outcome (locating an image by a search function is an outcome and visibility and searchability are affordances); that the proposed affordance has variability (features are binary and affordances have variability). So, using Evans et al.'s (2017) threshold criteria to distinguish affordances I find "true" potential digital archive affordances to analyze the archives with. These affordances are then real possible invitations for use that are relational constructs between the social and the technical.

Digital archive affordances through the lens of affordance theory are defined as the potential ways in which the archive-related possibilities and constraints associated with the material conditions and technological aspects of the digital archive are manifested in the archival characteristics of the studied digital archives. Identifying digital archival affordances is made through an overview of digital archive research. After identifying such affordances I examine manifestations of digital archive affordances in the studied archives.

The study of the political aspects of archives as the articulation of voice is a significant issue within the politics of archives and, with that, to feminist approaches within this field. Voice as participation is a theoretical model of power aspects as a participant-oriented process aiming to reveal how journalism research can explore, understand and critically discuss power aspects of archives by asking: whose voices participate where and with what consequences? This can be compared with a framing study of online news studying game frames or issue frames, and what is lacking or not. Framing aspects have been found through empirical studies in a way similar to affordances. I have developed an affordance theory approach for studying digital archives, where some affordances are there in varying degrees and some are not, and how this invites a critical discussion using voice and participation as power.

My affordance analysis moves from a descriptive to a more critical approach using Carpentier's (2016) model to critically analyze participatory media. The model articulates "layers" as fields and processes, making it possible to discuss and reason "how come" and "with what consequences"? Participation in this context refers to the equalization of power relations between privileged and non-privileged actors in formal

or informal decision-making processes. Real power is the ability to affect the outcome of such processes. This political approach shows that participation is an object of struggle and how ideologies defend certain participatory intensities. This makes it possible to discern problematic power discrepancies in power relations, by asking, for instance: What kinds of participation and power are present but also possible in digital women's archives of relevance to journalism history studies and digital journalism studies?

### Digital Archive Affordances

The following affordances have been identified in research on digital media and archives as being particularly relevant for digital archive studies:

1. Two key internet affordances: *hypertextuality* and *interactivity* (Wellman et al. 2003).
2. Two affordances specific to the potential of digital archives: *integration* and *customization*.
3. The affordance *visibility* as a possible action related to locating content.

*Hypertextuality* is associated with hyperlinks, which are seen as one of the most fundamental features of the web and as "intended connection[s] between segments of text" (Brügger 2017, 5). This includes an understanding of both the analogue and the digital. Analog segments of text were connected to each other earlier, and hyperlinks can exist on stand-alone computers as well as in local and global digital networks. I do not analyze hyperlinks as web data, but as ascribed affordances of the digital archive of hyperlinks: intended connections between segments of text.

*Interactivity* as an affordance refers to the degree of interaction with the archive. This includes all invitations to interact with the website, even for users to click on files as a form of co-creatorship that determines a record's meaning. Interactivity in a higher degree means user-contributed analysis and comment (Johnson 2017). Technical tools for interactivity are not only hyperlinks but also keyword searches, software downloads, as well as frequently asked questions (Aioki 2000). By interactivity, I also mean low degree interactivity, which makes it possible to discuss the digital of the archives at various degrees.

*Integration* relates to invitations to use digital archives for writing and researching in the same space (integration of parts of knowledge production). If the digital archive invites integration, the user can act as an "authoruser" in relation to material and also be invited to potential user collaborations. *Customization* is an affordance of digital archives that allows for the creation of personalized research spaces and classification systems. This means invitations for authorusers to assemble, upload, and save their own personalized collections of documents and material that they can describe and tag (Purdy 2011).

*Visibility* as an affordance refers to whether and how a piece of information can be located. Visibility leads to locating. In a way, this is the main affordance of an archive: to locate material. Visibility makes possible actions related to finding, confronting, viewing and consuming content. The search is a strong indication of visibility: "Visibility applies to any online technology that includes features to search for and find

information" (Evans et al. 2017, 43). Searchability as a function is something other than the visibility of search. Searchability relates to meta-data registration, search functions, to what actually is searched in, how searches are delimited, if they are in full text or via meta-data, and what the OCR quality is (Ben-David and Huurdeman 2014). An affordance perspective focuses on whether visibility is there or not and assesses it in terms of greater or lesser, or relative degree of visibility. Content visibility depends on a site's specific features, as well as the end user's application of specific features (Evans et al. 2017).

### **Voice as Participation to Analyze Power Matters**

What Couldry (2010) calls "new intensities of listening" I approach as "voice as participation." I do this by adapting and using Carpentier's (2016) model of a critical analysis of media participation and political agency. By translating voice that matters into voice as participation, it is possible to analytically discuss democratic implications of the digital archive affordances of the studied archives. Otherwise, the analysis would end in a simple description of how digital archives open up for more voices. With voice as participation, I am able to locate and discuss the participation of various voices, how they are made part of a context and made available for use, and how this says something about power in relation to journalism history studies.

Carpentier's model includes and integrates the participation process in its field by looking at how its actors make decisions and thereby express power. In this study, the model functions as themes for guiding the analysis. The first theme (process) is distinguishing how participation is located in particular archival processes, by asking: (1) In what way is an archive participatory or not, and what complexities are involved? (2) How are the processes situated within contexts that have an impact on them? The second theme (field) is focused on how the archives are part of a field or fields, by asking: How is the archive constructed and structured, with which knowledges, positions, interests, stakes, commodities and histories? The third theme (process and field) is analyzing the position of the archival processes in the field and how the relationships between the participatory processes and the field are organized. If we take women's voices in history-making as participation, then the participatory process takes place in certain ways within the field and across fields.

The fourth theme (actors) is focused on discerning actors that are active within the archival processes as well as the relations between these actors. The actors' identities and identifications are also considered. This means, for instance, contemplating how *Kvinnsam* is part of a gender studies and library organization and has a government mission to articulate women's history as a database. The digital is also considered an actor, as the materialization of trying to inform, and also invite to use and to link to other fields, shapes development. This also relates to whether the actors can be seen as privileged or not in the field. Furthermore, this makes it possible to discuss the degree to which, depending on field, the actors are not privileged. For example, mainstream journalism history and digital journalism do not deal with archives as a study object and gender issues.

The final theme is considering decision-making and power. This concerns what the decision-making moments are and what their significance is within the media





KvinnSam nyförvärv december 2017

Klicka på länkarna nedan för att se länestatus i GUNDA och beställa/köa.

Kvinn 000 17/7	Kleinert, Annemarie	Le "Journal des Dames et des Modes" ou la conquête de l'Europe féminine (1797-1839) / Annemarie Kleinert	2001
Kvinn 100 17/15	Boyle, Deborah A., author.	The well-ordered universe : the philosophy of Margaret Cavendish / Deborah Boyle.	2017
Kvinn 100 17/16	Claxton, Susanne, author.	Heidegger's gods : an ecofeminist perspective / Susanne Claxton.	2017
Kvinn 100 17/17	Kheel, Marti.	Nature ethics : an ecofeminist perspective / Marti Kheel.	2008
Kvinn 200 17/22		Female leaders in new religious movements / Inga Bardsen-Tellefsen, Christian Giudice editors.	2017
Kvinn 200 17/23		Unbinding Medea : interdisciplinary approaches to a classical myth from antiquity to the 21st century / edited by Heike Bartel and Anne Simon.	2010

**FIGURE 1**  
KvinnSam codes and categorization

process: How equal are the power relations in general when comparing the power position of the actors in each decision-making moment? What does an evaluation of the (un)balanced nature of the power positions of privileged and non-privileged actors show? This, for example, makes possible a discussion on how the decision-making moments for the archives within each archive and its institution are shaped by the voice of women in journalism history and women's history.

### The "Digital" of the Archives

WiJ displays hypertext through hyperlinks that allow users to continue clicking for more information. Hyperlinks mainly lead users through the project and to other information that concerns the Washington Press Foundation. One assigned link is to a YouTube interview, which is described as telling us that portions of the collection will be added to the WPCF website, and a text section describing the video with an invitation: "To view selected portions of the Oral History archives click here." Concerning interactivity, WiJ invites communication by addressing the user as a potential sponsor: "This work is made possible thanks to the generous support of our sponsors. If you would like to contribute to these efforts please send an email to the WPCF office." WiJ also invites users to interact through "Internships," which are related more to the Washington Press Foundation than to WiJ (<http://wpcf.org/internships/>). Still, the internships are described as providing "opportunities for women and minorities through internships at some of our nation's most prestigious media companies" (ibid).

KvinnSam displays hypertext by offering many options to click for more. KvinnSam users mainly start in navigations for the database (Figure 1). But *KvinnSam* also has interactivity in the form of invitations to contact KvinnSam generally and to suggest acquisitions specifically. It is the same document for suggesting acquisitions as for the library. In this information, it is the "Book" that can be purchased. As a user, one

The screenshot shows the Washington Press Club Foundation website. The header includes the logo and navigation links: HOME, ABOUT US, PROGRAMS AND EVENTS, WOMEN IN JOURNALISM, INTERNSHIPS, and ANNUAL CONGRESSIONAL. A sidebar on the left lists menu items: Archive Digitization, Project Overview, Highlights, Interviewees (selected), Interviewers, Repositories, Support and Funding, and Helen Thomas. The main content area is titled 'Women In Journalism | Interviewees' and features a search bar with the text 'Search All Interviews For:' and a 'Go' button. Below the search bar, a paragraph states: 'The interviewees fall into three professional generations:' followed by a bulleted list:
 

- women who began their careers prior to 1942;
- women who became journalists between the beginning of the World War II and the passage of the Civil Rights Act of 1964;
- and women whose careers developed after 1964.

 A subsequent paragraph explains the selection criteria for interviewees, focusing on diversity in race, locale, and type of journalism, as well as the woman's importance in her field and her impact on the industry and community. At the bottom of the section, there is a link for 'All Interviews A-Z'.

**FIGURE 2**  
Wij search

is guided toward mainly searching (in the top central menu) and acquisitions (in the lower right). The figure also shows in this section that it is the code for categorization that comes first.

The analysis shows that the archives do not manifest digital archive affordances to a great extent, a reasonably anticipated outcome. This reveals, however, an absence of the digital archival affordances of integration and customization. Concerning integration, *Wij* and *Kvinnsam* do not offer any facilities for the co-construction of meaning. The digital archives do not allow writing and research to occur together. Considering customization, *Wij* lacks the possibility of assembling and classifying, if you do not count the possibility to find all search words, as all words in the 10 interviews are searchable and appear as results in the hit list. One can download the material and customize it. But material is not readily accessible as an invitation to customization and offers for downloading are not integrated into the website. Customization also involves uploading, and *Wij* does not invite uploading of any material. In regard to *Kvinnsam's* customization, the site has a classification system. Customization is limited to saving a search, and uploading is not possible. The lack of integration and customization means that the voice of the authoruser (Purdy 2011) is lacking.

For journalism history studies, the archives offer trustworthy sources for writing women into history. Records are more or less accessible. At the same time, the archives do not fulfill any digital archive promises. The selection of the archives makes this hardly a surprise. What is gained through the analysis is an increased understanding of what digitized archives mean in relation to more digital archives, and what may be lost in a digitized archive where digital archive affordances are low or simply not there. The total lack of any digital archives within journalism to co-create indicates cultural

**FIGURE 3**  
Kvinnsam search

explanations. These can be related to the journalistic field where authority and autonomy are more valued than involving end-users in co-creation. Digital journalism is also a new field, starting in online news, mobile journalism and such, where the archive thus far is the Web as archive and online news archives. Within digital humanities, more experimental approaches to archives as well as to media history have been applied (Battershill et al 2017; Cambridge Digital Humanities 2018; The Center for Digital Humanities Princeton 2018). In Sweden, several other more digitally experimental archives have been initiated and developed, where the same main actors, KB and Libris, are involved. However, Kvinnsam is not part of these experiments. Another explanation can be that WiJ and Kvinnsam are not counter-archives. They are not experimenting to be accepted into established institutions. Both archives strive for legitimacy as well as to fulfill the aim of writing women into history. This is done by showing that highly competent women exist as well as including sources otherwise inaccessible (like feminist magazines and journals). Experimenting could mean losing legitimacy.

## Visibility

WiJ makes searchability visible in a not so visible way. Instead, the search is made part of the “interviewees” (Figure 2):

Kvinnsam places its search linking into the LIBRIS database and the field search that is the norm in database searches (Figure 3):

When performing a search, the WiJ archive provides results from only 10 interviews. Kvinnsam offers results from many more sources; for example, a search for “journalist” yields 41 hits. Visibility varies in these search results. WiJ gives instant

visibility to women journalists' reasoning on a particular chosen subject. Kvinnsam instead gives instant visibility to sources that need to be accessed and read through if one is to say something about women journalists. These digital archives only make visible portions of the archived material, not enough to constitute the entire research process. Both WiJ and Kvinnsam require the user to log in to get access to both actual and more material. But for journalism history studies, the digital archives in this form still provide visibility of more material that might not otherwise be considered research material.

None of the archives invite users to share or to contribute in any way to the voices in the archive (as was also mentioned in the former section). This is something typically valued by feminist archive initiatives. Instead, the archives invite users to engage in participation that is to listen to these chosen, catalogued and made available voices. Hence, voice as new intensities of listening is a possibility. In other digital archives, the actual sharing of material to contribute to an archive is more a matter of voice as expression. WiJ presents voices from the margins of journalism history to be listened to. Kvinnsam collects and make searchable central material for women's history studies that was previously too scattered to be easily accessed and/or used – making the material central in a collection where otherwise it would have been marginalized. This illustrates the value of distinguishing and analyzing voice as participation in relation to both voice as expression and voice in relation to listening.

### **Voice as Participation in WiJ and Kvinnsam**

The processes in focus here are aimed at creating and distributing (inform) an archive as an oral history archive of women journalists (WiJ) and a database for women's history (Kvinnsam). The fields are constructed and structured positions and interests, from both information management of knowledge production of history for journalism as a profession and of journalism history as knowledge production within women's studies. These differences invite different relationships between the participatory process and the field. When it is information management of history for journalism as a profession, it is the making of an oral history project within a journalist profession organization, distributing it digitally and through universities (WiJ). With respect to the latter, WiJ use repositories (universities), while Kvinnsam makes "woman" a starting point in material on journalists and journalism. And, information management of gender studies concerns how Kvinnsam is made part of the LIBRIS collection, organized through Secretariat for Gender Research.

The actors are the organizations behind the archive. For both WiJ and Kvinnsam, voice as participation implies that there should be relations between the archive and researchers and journalists. When analyzing the actors' identities and identifications, two aspects are displayed. WiJ's approach is that journalism is a profession where women have historically been and remain an integral component. Kvinnsam is a part of gender studies, library organization and with a government assignment to articulate women's history as a database. The digital as an actor is the materialization of trying to inform and also invite to link to other fields, shaping development. Both WiJ and Kvinnsam link into libraries and universities as repositories that shape their development and enhance their "voice as participation." Depending on field, the archives are

not privileged. Mainstream journalism history studies have always relied on archives, however, not dealing with archives as a study object, and women are downplayed. Both WiJ and Kvinnsam have non-privileged roles as both are peripheral to the central. At the same time, the archival agency is within each archive, separate from other powers of agency that shapes journalism both as a research field and as a profession. In women's studies, this separatist approach – being separate from dominant power – is both a necessity and a result of a power struggle within a field.

The politics made within each archive and its institutions are shaped by the power of the voice of women in journalism history and women's history. If we further this line of thinking to decision making within the various fields, the decision-making moments could mean all journalism history researchers should design based on an awareness of the digital archives of women's history that are available, as well as asking for and creating material that is lacking to do more inclusive journalism history.

### **Linking into Networks of Feminism**

By taking the form of an institution, the feminist project has historically gained power through archives. When an archive takes form as a, or within a, legitimate institution, the archive gains the power of being visible. One can express this as a formation where the collective voice as an institutional voice is crucial to becoming a public voice. Hence, voice as participation for the two archives relies on institutions to be a public voice.

WiJ and Kvinnsam illustrate this in various ways. The oral history field as actual voices is particularly evident in WiJ. It shows a materialization of history, or even a closeness to contemporary history, by providing archival material of interviews with the actual women pioneers of journalism. WiJ becomes a public voice for feminist journalism in embedding WiJ in a journalism foundation and the foundation's activities, like the aforementioned internships. The Swedish archive is more text-based and clearly expresses a multitude of material all in the public voices of "women." Kvinnsam becomes a public voice for women's history in general, linking into feminist networks by being a database in collaboration with the Secretariat for Gender Research.

Analyzing public voice as linking into networks of women's movements can also be understood as forms of resistance furthering a particular agency. Resistance is providing alternatives to a central, mainstream norm materialized as other voices, which implies that the archive remains an expression of voices, not as making a political impact. But in materializing as an archive, being placed within institutional contexts that (potentially) affect political matters, the oral histories and the data collection make expression political acts of resistance.

On another level agency and affecting political matters has to do with, who, then, is an end-user? Johnson (2017) points to the acute need in a digital archive for close engagement with end-users. Close engagement is within the feminist community trying to move from women being marginalized and silenced in the archives to being given voice. The agency is within these communities. But how many journalism history studies researchers and digital journalism researchers with an interest in women's journalism history know about these archives and/or use them? This type of end-user discussion is greatly needed.

## Discussion

WiJ and Kvinnsam imply that the democratic value of participation as agency in co-constructing history is severely impeded. At the same time, the digital archives do welcome the potential of voice as new intensities of listening by giving access to otherwise nonsearchable, nonvisible, and nonaccessible material of relevance to women's journalism history. Such potential depends on the power politics made visible in decision-making. This decision-making includes both actual contributions to digital archive material and actual uses of these archives. Public voice shows how the agency and political power of digital archives rely on institutions to be able to materialize as an archive and to (potentially) affect political matters and express political acts of resistance.

The archive as a service available to anyone implies participation as voice that includes new voices and voices with real agency. The two archives exemplify this. The archives try to make words survive as well as texts. Still, it is not the ordinary that is preserved in records and documentation in the archives. It is women, the voice of women, and it is the voice of exceptional and recognized women that are recorded, documented, categorized, and cataloged.

What archives and digital archives can do for journalism history research is to further build on women journalists and write them into history and also with more voices than those of privileged women. Peters (2008, 28) says "we live in a moment of acute archival sensibility, thanks in part to the internet." It is suggested this archival sensibility of journalism history research should be:

1. To use archives, digital archives and digital traces that acknowledges and analyzes a variety of materials.
2. To solicit, create, use and stimulate use of digital archives that include questions of rewriting journalism history.

For the politics of women's digital archives in journalism history, this means engaging in what archives exist, how they are used, and what value there is in various forms of web archives, digital archives and digital historiography. For digital journalism, this means a continuing focus on issues linked to digital archives as well as exploring what digital humanities and experimental archival studies means for developing the field. What archives and counter-archives exist and should exist? There could be experimental approaches, like The British Library Machine Learning Experiment (2018). There are also new and emerging approaches to the presentation of archival information online that show the potential for better and more connected archival information. There are several examples of critical archives — see, for instance, the list at Social Justice Digital Humanities Projects (2018).

The politics of women within digital archives for journalism history studies is present whether this is seen as central or peripheral to the field. This presence is essentially providing an argument for shifting perspectives of what centrality is and how the driving forces of history are being renegotiated in journalism history studies. To use simple and naïve understandings of a research field's strength as being mainly about what is "central," or what is usually done and with the usual material, is to shy away from truly democratic aspects of journalism history. It is to lend oneself to a hegemonic masculinity and consciously to both make invisible and to downplay important parts of

history, which, ultimately, continues a history peopled almost exclusively by kings and rarely by queens.

## ACKNOWLEDGEMENTS

I wish to thank the unnamed reviewers for appropriate and valuable comments and the guest-editor for providing guidance in relation to these recommendations. I also wish to thank Associate Professor Ann Werner (Linnaeus University, Kalmar) for constructive comments on conceptual clarity.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

## REFERENCES

- Aioki, Kumiko. 2000. "Taxonomy of Interactivity on the Web." <http://pdfs.semanticscholar.org/ed39/6d88541685ec738d1b9f5ac859e377036aa7.pdf>
- ArchiveGrid. 2018. "ArchiveGrid." <http://beta.worldcat.org/archivegrid/>
- Basu, Paul, and Ferdinand de Jong. 2016. "Utopian Archives, Decolonial Affordances: Introduction to Special Issue." *Social Anthropology* 24: 5–19.
- Battershill, Claire, Helen Southworth, Alice Staveley, Michael Widner, Elizabeth Willson Gordon, and Nicola Wilson. 2017. "Introduction." In *Scholarly Adventures in Digital Humanities Making The Modernist Archives Publishing Project*, edited by Claire Battershill, Helen Southworth, Alice Staveley, Michael Widner, Elizabeth Willson Gordon, and Nicola Wilson. Cham: Springer.
- Beasley, Maurine. 2010. "Recent Directions for the Study of Women's History in American Journalism." *Journalism Studies* 2: 207–220. doi:10.1080/14616700117394.
- Beasley, Maurine. 2015. "Women in Journalism: Washington Press Club Foundation Oral History Archive." *American Journalism* doi:10.1080/08821127.2015.999637
- Ben-David Anat, and Hugo Huurdeman. 2014. "Web Archive Search as Research: Methodological and Theoretical Implications." *Alexandria* 25 (1–2): 93–111.
- Brügger, Niels. 2017. "Connecting Textual Segments: A Brief History of the Web Hyperlink." In *Web 25: Histories from the First 25 Years of the World Wide Web*, edited by Niels Brügger, 3–28. New York: Peter Lang.
- Burton, Antoinette. 2003. *Dwelling in the Archive: Women Writing House, Home, and History in Late Colonial India*. New York: Oxford UP.
- Cambridge Digital Humanities. 2018. <http://www.cdh.cam.ac.uk/research/research-impact-1>
- Carpentier, Nico. 2016. "Beyond the Ladder of Participation: An Analytical Toolkit for the Critical Analysis of Participatory Media Processes." *Javnost – The Public* 23 (1): 70–88.
- Casswell, Michelle. 2014. *Archiving the Unspeakable. Silence, Memory and the Photographic Record in Cambodia*. Madison, Wisconsin: The University of Wisconsin Press.
- Centre d'Archives et de Recherches pour l'Histoire des Femmes. 2018. [http://www.avg-carhif.be/cms/sites\\_pays\\_en.php](http://www.avg-carhif.be/cms/sites_pays_en.php)

- Cifor, Marika, and Stacy Wood. 2017. "Critical Feminism in the Archives." In *Critical Archival Studies*, edited by Michelle Caswell, Ricardo Punzalan, and T-Kay Sangwand. Special issue, *Journal of Critical Library and Information Studies* 1 (2).
- Couldry, Nick. 2010. *Why Voice Matters: Culture and Politics After Neoliberalism*. London: Sage.
- Derrida, Jaques. 1996. *Archive Fever: A Freudian Impression*. Chicago: University of Chicago Press.
- Djerf-Pierre, Monika, Marina Ghersetti, and Ulrika Hedman. 2016. "Appropriating Social Media." *Digital Journalism* 4 (7): 849–860.
- Eichhorn, Kate. 2014. *The Archival Turn in Feminism: Outrage in Order*. Temple UP.
- Evans, Sandra, Katy Pearce, Jessica Vitak, and Jeffrey Treem. 2017. "Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research." *Journal of Computer-Mediated Communication* 22: 35–52.
- Fuchs, Penny. 2003. "Journalism History." *Athens* 28 (4): 191–196.
- Gauthereau, Lorena. 2017. "Incubator: Decolonizing the Digital Humanities." *Recovering the U.S. Hispanic Literary Heritage Blog*, November 20. <http://recoveryprojectappblog.wordpress.com/2017/11/20/incubator-decolonizing-the-digital-humanities/>
- Gibson, James. 1979/1986. *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Erlbaum.
- Hoskins, Andrew. 2017. "The Restless Past. An Introduction to Digital Memory and Media." In *Digital Memory Studies: Media Pasts in Transition*, edited by Andrew Hoskins. New York: Routledge.
- Howell, Chuck. 2006. "Dealing with Archive Records." In *Methods of Historical Analysis in Electronic Media*, edited by Donald Godfrey, 305–348. Mahwah, NJ: Erlbaum.
- Johnson, Valerie. 2017. "Solutions to the Silence." In *The Silence of the Archive*, edited by David Thomas, Simon Fowler, and Valerie Johnson, 141–162. London: Facet.
- Kvinnsam. 2018. <http://www.ub.gu.se/kvinn/Kvinnsam/>
- Lubelski, Sarah. 2014. "Kicking Off the Women's "Archives Party": The World Center for Women's Archives and the Foundations of Feminist Historiography and Women's Archives." *Archivaria* 78.
- Nagy, Peter, and Neff, Gina. 2015. "Imagined Affordance: Reconstructing a Keyword for Communication Theory." *Social Media + Society* 1: 1–9.
- Nerone, John. 2011. "Does Journalism History Matter?" *American Journalism* 28 (4): 7–27.
- Ney, Birgitta. 2006. *Kvinnosaken i Pressen - Kvinnor, Journalister och Tidningstexter*. Stockholm: Institutionen för etnologi, religionshistoria och genusstudier.
- Peters, John Durham. 2008. "History as a communication problem." In *Explorations in Communication and History*, edited by Barbie Zelizer. Oxon: Routledge.
- Purdy, James. 2011. "Three Gifts of Digital Archives." *Journal of Literacy and Technology* 12 (3): 24–29.
- Robertson, Craig (editor). 2011. *Media History and the Archive*. London: Routledge.
- Rogers, Richard. 2017. "Doing Web history with the Internet Archive: Screencast Documentaries." *Internet Histories* 1 (1–2): 160–172.
- Shaw, Adrienne. 2017. "Encoding and Decoding Affordances: Stuart Hall and Interactive Media Technologies." *Media, Culture & Society* 39 (4): 592–602.
- Social Justice Digital Humanities Projects. 2018. <http://criticaldh.roopikarisam.com/social-justice-digital-humanities-projects/>



- Steiner, Linda. 2017. "Gender and Journalism" In *Oxford Research Encyclopedia of Communication*.
- The 1947 Partition Archive. 2018. <http://www.1947partitionarchive.org>
- The British Library Machine Learning Experiment. 2018. <http://blbigdata.herokuapp.com/>
- The Center for Digital Humanities Princeton. 2018. <http://cdh.princeton.edu/projects/>
- The Herstory: JAWS Oral History Project. 2018. <http://herstory.rjionline.org/001.html>
- The Interviews: An Oral History of Television! 2018. <http://interviews.televisionacademy.com/>
- The Journalism Digital News Archive. 2018. <http://www.rjionline.org/stories/series/journalism-digital-news-archive>
- The Women Who Rock Digital Oral History Archive. 2018. <http://content.lib.washington.edu/wwwweb/>
- Theimer, Kate. 2012. "Archives in Context and as Context." *Journal of Digital Humanities* 1 (2).
- Tenenboim-Wenblatt, Keren, and Motti Neiger. 2018. "Temporal Affordances in the News." *Journalism* 19 (1): 37–55.
- Thomas, David, Simon Fowler, and Valerie Johnson. 2017. *The Silence of the Archive*. London: Facet Publishing.
- Tusan, Michelle. 2005. *Women Making News: Gender and Journalism in Modern Britain*. Urbana, IL: UI Press.
- Uricchio, William. 2014. "History and Its Shadow: Thinking About the Contours of Absence in the Construction of Media History." *Screen* 55 (1): 119–127.
- Wellman, Barry, Anabel Quan-Hasse, Jeffrey Boase, Wenhong Chen, Keith Hampton, Isabel Diaz, and Kakuko Miyata. 2003. "The Social Affordances of the Internet for Networked Individualism." *Journal of Computer Mediated Communication* 8(3).
- Whitt, Jan. 2008. *Women in American Journalism: A New History*. Urbana and Chicago: UI Press.
- Wikipedia. 2018. "Wikipedia List of Archives." [http://en.wikipedia.org/wiki/List\\_of\\_archives](http://en.wikipedia.org/wiki/List_of_archives)
- Women in Journalism. 2018. <http://www.wpcf.org/women-in-journalism/>
- Zanish-Belcher, Tanya and Kären Mason. 2007. "Raising the Archival Consciousness: How Women's Archives Challenge Traditional Approaches to Collecting and Use, Or, What's in a Name?" *Library Trends*. doi:10.1353/lib.2008.0003.

# DIGITAL ARCHIVING AS SOCIAL PROTEST

## *Dalit Camera* and the mobilization of India's "Untouchables"

**Subin Paul and David O. Dowling**

*The relationship between journalism and social marginalization is a relatively understudied area in digital journalism studies. Our case study of Dalit Camera (DC), an online news archive and chronicle based in India, examines how historically disadvantaged Dalits, or "Untouchables," are leveraging digital tools to narrate their oppressive past to the outside world parallel to the rise of political censorship in India. As part of its archiving process, DC is preserving footage of Dalit resistance against hegemonic domination by caste Hindus. Through their grassroots network of citizen journalists, DC is also engaged in reporting caste-based discrimination and violence today, contributing to the Dalit social movement for equality and justice. Our study provides the first examination of Dalit social protest as a function of digital news archiving, in the process bringing a non-Western subject typically reserved for Subaltern Studies to digital journalism studies as a potent example of citizen journalism and participatory online culture in a censorious media climate. We argue that the growing field of digital journalism studies must leverage productively with area studies scholarship, such as Subaltern Studies, in order to advance a deeper understanding of journalism cultures in the Global South.*

### **Introduction**

After enduring a brutal beating at the hands of members of a Hindu nationalist student organization, Ravichandran Bathran, a former doctoral student at the English and Foreign Language University in Hyderabad, resolved to establish an online archive to document the violence perpetrated against the outcastes of India known as Dalits or "Untouchables." Being a Dalit himself and belonging to the manual scavenging community—which cleans toilets that do not have a modern flush system—Bathran, like other Dalits throughout the country, was a witness to and victim of caste oppression. While caste discrimination has been a staple of Indian society for centuries, the specific manifestations of oppression rarely make headlines in the mainstream news media, which are owned and operated primarily by upper-caste Hindus (Rao and Wasserman 2015; Mody 2015). It is within this context that Bathran turned to a relatively democratic space of the internet to launch an archive called *Dalit Camera (DC)* in 2012.

*DC* is an archive in the form of a YouTube channel and a website (dalitcamera.com) that records news, narratives, public meetings, songs, talks, and discussions on Dalits and other underrepresented sections of the Indian society. Its stated purpose is to document views from activists, students, and intellectuals of the Dalit community primarily as “a response and counter-view to the cartelized hegemony of the English news” (Bhim 2018). *DC* has thus positioned itself in opposition to the “Murdochization” of India’s mainstream media associated with rampant corporate corruption (Sonwalkar 2017). The website functions mainly as a “textual” counterpart to the YouTube channel, which on an average is updated daily. It is through this close monitoring of the present and the continuous (although uneven) accumulation of online items that news websites, such as *DC*, become archives (Bødker 2017). In other words, *DC* can be understood as an archive because news sites, through accumulation, inadvertently become archives in the sense that news sites are a “combination of chronicling and archiving” (59). Furthermore, *DC*’s “interactive affordance” (Bødker and Brügger 2018) facilitates digital citizen journalism—the act of ordinary people creating online media content that includes news and information (Wall 2015). *DC*’s mission to “document perspectives on/voices of Dalits” therefore casts its contributors as both citizen journalists and archival documentarians. The contributors work in collaboration with *DC*’s volunteers, who are mainly students based at the metropolitan cities of Hyderabad, Mumbai, and Kolkata, maintain sole access to upload materials to the website and channel.

In this examination of the content and practice of digital archiving and chronicling as social protest, we approach archives according to Weld (2014) not only as sources of data to be excavated by researchers but also “as more than the sum of their parts—as instruments of political action, implements of state formation (‘technologies of rule’), institutions of liberal democratization, enablers of gaze and desire, and sites of social struggle” (13). Instead of mere “institutions of memory” (Hartley 2012, 157) regarded in a purely technicist manner, archiving in our study is conceived as a social and political practice (Coudry 2008; Udupa 2016).

Informed by the work of Castells (2015) on how various social movements across the world have benefited from digital technologies, we evaluate the potential and limitations of an online archive such as *DC* in terms of social justice for and political upheaval of Dalits and other subaltern (i.e., disenfranchised and repressed) groups to answer the following questions. How does *DC* represent Dalit issues and voices in India? Does *DC*, both as a digital archive and as a participatory chronicle, contribute to the Dalit social and political movement for equality and justice, and if so, in what ways? What kinds of problems does *DC* encounter in representing Dalits and other marginalized groups? And lastly, how can the *DC* archive advance subaltern voices and (re)construct journalism history from the perspective of common people rather than elites of South Asia?

Our goal in this article is to explore these questions through a discussion of exemplary cases recorded in *DC*. In doing so, we are inspired by Pandian’s (2008) recommendation to produce “enabling re-descriptions of life-worlds” and thereby facilitate “morally and politically enabling knowledge(s) about Dalits and other subaltern groups” (34). This essay begins by situating Dalits and their media within their larger sociological setting. It then analyzes the work and function of the *DC* archive, specifically its YouTube channel and blog, parallel to theoretical insight on social movements in

digital space. In the next section, we focus on how the *DC* archive helped popularize a Dalit social movement in the wake of a key student-activist's suicide. The conclusion then addresses the limitations, possibilities, and implications of the *DC* archive for Dalit political activism and digital journalism studies.

### Situating Dalits and Their Media

Despite the implementation of democratic ideals such as equality and universal suffrage in the 1950s, Dalits remain at the bottom of the caste system—the principal mode of social stratification in India (Rao and Mudgal 2015). In colonial India, various names were applied to the castes that the British combined under the label “Untouchables.” Scheduled Caste (SC) is the official, administrative term, derived from the list, or “schedule,” in the 1935 constitution that recorded such castes (Jeffrey 2001). Dalit is the favored term, though, meaning “oppressed” or “ground down.” To validate exclusion of Dalits from information, ancient Hindu scripture is sometimes called on: “If a Sudra [and Dalits were even lower in status] ... listens to a Vedic recitation, his ears shall be filled with molten tin or lac; if he repeats it, his tongue shall be cut off; if he commits it to memory, his body shall be split asunder” (*Gautama Dharmasutra* 1999).

Although the Indian constitution has formally abolished untouchability, in practice, it still persists in the country, especially in rural areas. Dalit movements for social equality are restricted, though in some states such as Uttar Pradesh, Dalits have become an increasingly organized political force during the past few decades (Belair-Gagnon et al. 2014). Historically, Dalits were forced to ask permission from landlords before leaving their villages or conducting marriages, and they faced beatings if they transgressed. As a population, Dalits are predominantly poor, heavily discriminated against, overwhelmingly rural and more than half illiterate (Jeffrey and Doron 2012). Social welfare measures such as affirmative action, in which a proportion of seats in government jobs and higher education are reserved for Dalits, have produced limited results. Consequently, Dalits who comprise approximately 17 percent (nearly 200 million) of India's population continue to be socially stigmatized.

In electoral politics after India's independence in 1947, Dalits were often treated as additional voters for the landlords on whose land they lived and worked. Dalit political movements were limited with the only “major” party being the state-level Bahujan Samaj Party (BSP), which won the 2007 elections in Uttar Pradesh, but has hitherto failed to have national influence (Jeffrey and Doron 2012). Apart from the BSP, political parties from the left have forged, to various extents, alliances with Dalits. Such overtures, however, have frequently met with criticisms from a section of Dalit activists and intellectuals (Mishra 1999).

Dalit-produced media originally appeared in the early twentieth century with *Achut* (“Untouchable”), the first notable Dalit newspaper published in 1917 at Delhi. Around the same time, Ayothee Das, a Dalit leader from the Madras Presidency, started a newspaper called *Oru Naiya Piasa* (“One New Paisa”) (Pandian 2007). Since then, there were several Dalit newspapers and magazines published from cities, towns, and regions across the country. Yet, their number was negligible in comparison to mainstream publications. Furthermore, as Jeffrey (2001) noted, all Dalit publications were fringe, often with a literary emphasis and little influence beyond the circle that produced them.

Language divisions within India, and caste divisions among Dalits themselves, constricted the market for publications trying to cater to Dalit interests. Lacking access to traditional forms of media, Dalits developed their own limited media and cultural practices with which they shaped their identities. Over the past two generations, a tiny Dalit middle-class emerged which produced internationally recognized Dalit literature of autobiography and poetry in English and several Indian languages (Thirumal and Tartakov 2011). Similarly, film scholars noted an interest in Dalit and caste issues in South Indian cinema. Recently several visual documentaries were commissioned by government and private agencies as ethnographic films on Dalits. These documentaries traced their origins to the work of filmmakers such as Anand Patwardhan.

Dalits rarely worked in mainstream news media till the twentieth century because of a combination of structural and ideological bars, which made it immensely difficult for Dalits to enter or advance in the journalism industry. Jeffrey (2001) claimed that the position of Dalits in Indian newsrooms was at least two generations removed from that of Blacks in American newspapers. Partly owing to the near absence of Dalits in newsrooms, mainstream media typically refuse to give prominence to atrocities suffered by Dalits and systematically deny grassroots Dalit movements space in publications. The Dalit voice was missing from or marginal in most alternative and citizen journalism outlets that were established after the internet boom in the 2000s (Poell and Rajagopalan 2015; Chadha and Steiner 2015).

Yet, with the advent of the internet, there was never before as much opportunity for dialogue outside of mainstream media where the marginal can not only speak, but can also expect a response (Mitra 2001). Building on this opportunity, Dalit internet groups such as *ambedkar.org* (named after the Dalit iconic figure and architect of the Indian Constitution B.R. Ambedkar) and *dalitstan.org* (which since 2006 has been blocked by the Indian government) shaped crucial issues in the construction of an alternate Dalit history and identity (Thirumal 2008). As Dalits are present throughout the country, but nowhere in majority, they constitute perhaps the only community other than the Brahmins to display an eagerness to share a pan-Indian identity. This eagerness, Thirumal (2008) argued, can be realized through the use of online media by the Dalit middle-class. Chopra (2006), however, cautioned that in its online participation, Dalit discourse may tend to mirror the dominant mode of representation by caste Hindus that can stake a claim to cultural ownership of the nation, even as Dalits remain opposed to the ideology of Hindu nationalism. The voice of Dalits received an online fillip with the establishment of discussion portals such as *Round Table India* and *Ambedkar's Caravan* in 2008 and 2009. The group's online community further expanded with *Savari*, a website featuring the writings of Dalit, tribal, and Bahujan ("peoples in majority") women. Taken together, these portals showcase issues that are otherwise sidelined or misrepresented in mainstream news media.

DC joined this emerging Dalit online space in 2008 when it posted the video of a Dalit Panchayat ("assembly") leader in the southern Indian state of Tamil Nadu who was attacked by caste Hindus. Since then, DC focused on a range of issues relating to Dalits, Muslims, women, and other underrepresented groups. "As a student, I didn't have the means to start a newspaper or television channel [but I could] film instances of discrimination," said Bathran, adding, "When we hear of an atrocity, we interview the victim, put up whatever raw footage we have, record dalits' opinion and upload

the video" (Dhillon 2014). Most of *DC*'s contributors are students with little to no formal journalism training. Beside financial donations, the work of *DC* is self-funded. In terms of audience, *DC*'s YouTube channel is still nascent with about 20,000 subscribers. The channel is arguably far more influential than that number suggests because Dalit portals including *Round Table India* frequently republish *DC*'s posts on their websites.

The significance of *DC*'s function as a progressive alternative to mainstream media intensified in the wake of the 2014 elections that brought the Bharatiya Janata Party (BJP) to power. Since these elections, which led to Narendra Modi's appointment as Prime Minister, a new wave of press censorship has taken over India, driving out editors and journalists critical of him, and ushering in a new era of intolerance for anything other than lapdog journalism. Dalits and their agitation for caste annihilation clearly had no place in the newsroom and headlines given this convergence of State and media under "the expectation that the news media are called upon to act as cheerleaders of the state instead of holding it to account" (Sonwalkar 2017, 535).

*DC*'s founder and contributors see themselves as "archivists" for they do not "add value to the protests [or events] on the ground" (Aravind 2016). *DC*'s archiving practices nonetheless differ from traditional archiving in that the former is a more dynamic endeavor—filled with timely ripostes and commentaries against opposing mainstream narratives—than an exercise confined to designing and maintaining web portals for information storage and display. Its function thus far exceeds the mere "curation of data" (Udupa 2016). In this dynamic political activity, archiving becomes a "living library" not "stuffed into [static] libraries or state vaults," but a struggle over archiving power as author-function to "make and command what took place here or there, in this or that place, and thus what has an authoritative place in the contemporary organization of social life" (Povinelli 2011, 150–152). Moreover, past news in an online news site and archive, such as *DC*, accumulate and increasingly become searchable, by "dragging along their own genealogies" (Bødker 2017, 59; Bødker and Brügger 2018). Such an "archive," therefore, is a useful resource not only to future journalism scholars aspiring to write an alternative history from the perspective of the oppressed, but more directly to the very subaltern groups which are looking for new political paths and movements ahead.

### **Social Movements in the Digital Age**

According to Castells (2015), social movements have transformed profoundly in the digital age, as technologies have enabled "the rise of new forms of social transformation" (47). In particular, the pattern he identified in social movements originating on the internet and expanding into urban spaces describe a similar one to that of the Dalit protests that leverage *DC* on YouTube. Castells (2015, 46) noted that in Tunisia and Iceland, "the movement went from cyberspace to urban space, with the occupation of the symbolic public square as material support for both debates and protests." *DC* has similarly leveraged online media to publicize its protests on behalf of Dalits. Not only does *DC* function as a clearinghouse of video documentation of abuses to this minority population, it also has drawn protesters together. In the wake of a middle-class 23-year old medical student's gang rape by twenty men in 2012, *DC* filmed a video of a Dalit activist Rekha Raj voicing criticism in a public setting. Her point was

that widespread rape of Dalit lower-caste women in farming at the hands of upper-caste landowning men had received no attention in comparison to the middle-class victim. The protest highlighted how chronic rape is perpetuated by the prohibition of free speech among Dalits, which in effect silences victims (Mehta 2014). *DC* thus not only provides a voice for such protesters and an archive of atrocities suffered by Dalits, it also functions as what Castells called “a hybrid public space made of digital social networks,” which for *DC* consists of subscribers and followers via social media on YouTube. For this grassroots uprising, “a newly created urban community is at the heart of the movement, both as a tool for self-reflection and a statement of people’s power” (Castells 2015, 46).

Much of the political efficacy of *DC* lies in its status as a YouTube channel, whose medium of visual communication features raw video, in some cases edited in documentary formats by mostly untrained videographers. The rawness of the footage creates authenticity as well as intimacy, both of which play a key role in rallying support for Dalit demonstrations. Similarly, the Egyptian revolution relied heavily on YouTube to bring a deeply human element of lived experience to the event that mobilized resistance, which was further coordinated through Twitter and other SMS. “Videos of security forces treating the protesters brutally were shared via the internet,” Castells observed, “exposing the violence of the regime in unedited form.” Just as several key videos on *DC* have been viewed more than 50,000 times despite the channel having about 20,000 subscribers, the Egyptian revolution depended on the circulation of compelling raw videos through social media. “The viral nature of these videos and the volume and speed with which news on the events in Egypt became available to the wider public in the country and in the world was key to the process of mobilization against Mubarak” (Castells 2015, 60).

As with the Egyptian revolution’s use of multimodal autonomous communication to break through the wall of silence and empower the oppressed to voice their dissent, the previously silenced Dalits have come forward with their testimony on *DC*. The digital archive in this case has a similar effect of breaking “the barriers of isolation, making it possible to overcome fear by the act of joining and sharing” (60). Since Dalits are denied access to local public platforms for speaking out and are systematically erased from public view by news organizations that refuse to cover their stories and hire them as journalists, the lack of a medium of communication kept them in isolation thus disabling organization in protest. With *DC*, women Dalit farm-workers can view and share each others’ stories of sexual violence and gather in unprecedented numbers in public spaces to voice their dissent. Linkan Subudhi epitomizes this digital generation of Dalits newly empowered by the digital archive to advocate on behalf of women rape victims and abused children. “Caste is not a reason for being raped,” she said in one video. “Any woman is unsafe” (Mehta 2014).

Social media platforms such as Twitter that offer synchronous or near synchronous communication have been effective in coordinating activist organization and protests (Shirky 2009; Castells 2015). This is due to the broad reach and frictionless replicability of SMS. YouTube, by comparison, functions as an asynchronous platform “to offer people access to like-minded others and support, whether those others are online simultaneously or not.” In this manner, *DC* would not be functional, nor as politically potent, if its material were restricted to an SMS such as Twitter. But on

YouTube, *DC* can perform a digital archiving function as a repository of the Dalit plight, from raw footage of their victimization to public protests, while also providing a means of social integration and network support. Such support “enables people to feel part of a group whose members have common interests and concerns,” as with *DC* (Cutrona and Russell 1990, 322). Whereas YouTube originated as a grassroots online community for the amusement of amateur videographers, hobbyists, and later increasingly entrepreneurial vloggers, its digital affordances as asynchronous communication have been reinvented as a tool of political resistance.

### Channeling the “Chalo HCU” Movement Through Digital Archive

As stated earlier in this essay, we view the *DC* archive not as a “static object,” but as a “dynamic and virtual concept” (Cook 2001) or entity, which is constructed socially and culturally through participation. Conceiving archive in such a fashion allows us to be attentive to the practice of archiving itself; one in which record-making and record-keeping may involve political negotiations between the state and socially mobilized groups. For example, the kind of content included and excluded from the Dalit archive bear the imprint of the struggle and violence of the colonial past (Spivak 1985) and political present. Furthermore, influenced by Foucauldian and Saidian reflections on knowledge production, scholars of Subaltern Studies—a field within area studies scholarship that is interested in understanding postcolonial societies from a bottom up perspective—have alerted us to imagine archive not “as a store of transparent sources but as a veritable site of power” (Bandyopadhyay 2011b). This study is thus less concerned with mining or establishing the “truth” in the digital archive (Stoler 2002) than in understanding the cultural and political conditions and processes of the production of that “truth” as well as its implications for the larger Dalit community and movement.

This critical perspective lends insight into a “movement” called Chalo HCU (“March to HCU”). Chalo HCU started because a doctoral student, Rohith Vemula, committed suicide by hanging himself on 17 January 2016 after he was subjected to caste oppression in Hyderabad Central University (HCU), a premier social science institute in India. As an active member of the Ambedkar Students Association, Vemula was earlier expelled from HCU because of his alleged involvement in threatening the leader of the Akhil Bharatiya Vidyarthi Parishad (ABVP), a Hindu nationalist student organization. With ABVP being the protégé of the ruling right-wing BJP, there was considerable pressure on the HCU administration and state government to water down the protest that followed Vemula’s suicide, including creating a dominant, disputable narrative that Vemula was not a Dalit (Apparasu 2017; Prasad 2017). On the other hand, anti-caste groups including Dalits termed Vemula’s death an “institutional murder” (PTI 2016) and organized a protest against the HCU administration.

While mainstream media’s initial response to the protest was at best tepid, *DC* provided a blow-by-blow coverage of the protest as its volunteer Dharmateja LV filmed the protests which were then uploaded to *DC*’s YouTube channel (Nagpaul 2017). Dharmateja’s videos presented the first eyewitness, detailed accounts of the protest. These unedited videos, which included testimonies of Dalit students speaking about caste discrimination they faced, were instrumental in publicizing the protest to the wider population in India and abroad. Students at several Indian universities and



institutes including the Tata Institute of Social Sciences demanded justice for Vemula. In addition, Dalit organizations and Indian diaspora organized protests in places such as London, San Francisco, and Johannesburg (Khan 2016). *DC* contributor Dharmateja's background—a computer engineer with no formal journalism experience—demonstrates how ordinary, motivated citizens armed with technology can engage in digital archiving and citizen journalism. In all, *DC* uploaded more than 400 live videos relating to protests on Vemula's suicide on YouTube. *DC*'s archiving in the "Chalo HCU" protest, as in other instances, have benefited from the massive growth of digital media and innovations in recording and storage devices in the past two decades, which is shaped in part by informational economy's voracious appetite for data as a source and medium for capital accumulation (Harvey 2010) and peer-driven technological innovations that have changed the scale, architecture, and experience of archiving by adding speed (Udupa 2016).

In addition to the videos on YouTube, *DC* spearheaded the movement for Vemula's justice by publishing commentaries, interviews, speeches, memorandums, press releases, letters, songs, and digital installations by intellectuals, artists, and activists on its blog. One of *DC*'s artists paid tribute to Vemula through a digital slideshow. With Vemula's photograph inscribed in a star in the backdrop, the slides contained agentic quotes such as "State should not have the right to hang someone. Especially when the state is Brahmanical" and "When religious instruction comes with a proxy term, 'CULTURE', the resistance should be registered strongly" (Shobana 2016). *DC*'s volunteers also transcribed many of the videos and interviews into English for maximizing the reach of the posts. "Rohith Vemula lives as long as the caste system lives in this country, as long as oppression continues in the name of caste in this country," said the poet Bojja Tharakam (2017) in an interview with *DC*. In a speech by Vemula's mother, Radhika Vemula (2017) urged Dalits, Muslims, Adivasis, Bahujans, Women and Communists to unite and build political coalitions to defeat BJP in elections and capture power at the center. This speech as well as the larger movement indeed drew the otherwise internecine political groups together to demand justice for Vemula in a more concerted and public manner. Thus, an eclectic range of journalistic, oratorical, literary, visual and sonic representations of Dalit resistance on *DC*'s archival space—which were then shared through e-mail, Twitter (#ChaloHCU and #RohithVemula), Facebook, placards, and word-of-mouth—created an ephemeral environment of proximity for Dalits across the country and internationally. Such representations asserted, as Matzner (2014) noted in her study of Dalit sonic cultures, claims to Dalit belonging and presence in online space to outsiders for whom Dalits' physical presence is considered defiling.

Such archiving from the bottom, or "reverse archiving" (Bandyopadhyay 2011a), has several elements in common with citizen journalism. These include emphasis on issue-based rather than profit-driven news, a focus away from traditional (often governmental) newsmakers to marginalized laypersons, and a more egalitarian network of news contributors who are mostly ordinary people focused on their immediate communities (Campbell 2015; Meadows 2013). Because the majority of Dalits reside in rural India, mainstream media mostly based in urban regions cite issues of "proximity" and "staff crunch" as reasons for the lack of news on Dalits (Dash 2013). Additionally, the growing corporatization and political partisanship of mainstream media provide an

impetus for grassroots, citizen-driven journalism (Paul 2017; Sonwalkar 2017). Building on the cultural and technological shift, grassroots journalism with its network of local activists and intellectuals is able to contribute to *DC*, which thrives on input from these people. Consequently, the assemblage of news from contributors at the grassroots level results in an archive that is entirely based on marginal voices. This represents a significant democratic breakthrough given that such hyperlocal news stories and narratives were almost impossible to be recorded and accessed at a single location across spatial and temporal boundaries before the advent of digital technologies.

Citizen journalism is consonant with *DC*'s journalistic roles. Hanitzsch (2011) identifies four types of roles for journalists: populist disseminators, detached watchdogs, critical change agents, and opportunist facilitators. Whereas *DC* certainly fulfills the role of critical change agent on behalf of the Dalit movement, its watchdog commitment to advocacy journalism is hardly detached, and it certainly harbors no opportunistic mercenary traits of profit seeking, as evidenced by the largely untrained amateur staff of citizen contributors. It is a grassroots platform, and thus reflects a populist fervor, although one representing a minority marginalized population of Dalits not to be confused with Modi's mainstream majoritarian brand of populism. Furthermore, *DC*'s journalistic culture exhibits elements of resistance to the notion of objectivity (especially as ethical neutrality) according to Agarwal and Barthel (2015, 381) by its pursuit of fairness through advocacy. The platform, however, does not conform to new online norms of individualism influencing Western newsrooms, particularly in the United States, but instead lays collective emphasis on the transparency and risk taking common in digital culture.

*DC* not only presents a potent tool for Dalits to organize socially and politically against caste oppression, but also provides a rich archive of indigenous perspectives and journalistic narratives from the peripheries. Journalism thus forms the base for continuity and durability of communities "by arresting time in various forms of texts, which then inform collective memory and which accumulate as cultural heritage and material for the writing of history" (Bødker 2017, 57). The peripheral accounts could therefore be immensely useful to future journalism historians who are sensitive to marginalized accounts and are looking to do what Thompson (1966) famously phrased "history from below." In this scholarly endeavor to (re)interpret the present with the benefit of hindsight, a Dalit archive and chronicle such as *DC* would be instrumental in opening new avenues for the field of journalism history, which has hitherto focused predominantly on Western societies or their English-language, mainstream counterparts in the Global South.

## Conclusion

In her article on expanding citizen journalism scholarship, Wall (2015, 806) asks researchers to examine understudied areas, such as "the ways citizen journalism intersects with *race, gender, class, and other categories of marginalization*." Our study of *DC* is among the earliest attempts to analyze the intersection of citizen journalism and the sociological category of caste. Furthermore, while the dominant theories used to explore citizen journalism are the sociology of journalism, the public sphere, and convergence cultures (Borger et al. 2016), we have demonstrated the utility of an

interdisciplinary approach, one that weds scholarship from digital journalism studies with literature on archival science, social movements, and Subaltern Studies. We argue that the growing field of digital journalism research must leverage more productively with area studies scholarship, such as Subaltern Studies, to produce new knowledge about journalism cultures in the Global South context.

Such an interdisciplinary approach honors the diverse digital literacies and production practices of the twenty-first century, one particularly effective for culturally and industrially situating non-Western digital media such as *DC*. Although new automated tools for analyzing journalistic texts have emerged as detailed in *Digital Journalism's* special issue on "Research Methods in an Age of Digital Journalism," their development mainly by United States and European scholars raises concerns about their applicability to *DC* and other non-Western digital media driving social movements of oppressed populations. As Bourmans and Trilling (2016) warn, although computational approaches to digital journalism studies can provide key insight into the liquid nature of online journalism, their use demands a deep understanding of the subject and media content first and foremost, a point particularly significant in the case of Subaltern media. The study of liquidity, which is essential to the understanding of the digital archive, "cannot be based solely on random samples, simply because they often presuppose more specific strategic samples," such as those used to examine the archive as political protest in *DC* (Widholm 2016, 28). The liquid nature of *DC* we have located in its platform convergence between its YouTube channel and website, complex "external" connectivity through social media, and escalating audience interaction (Deuze 2008), the latter of which is a potent source for citizen journalism. Its archival function, however, explicitly resists textual ephemerality associated with liquid media. All documentary evidence and testimony published on behalf of Dalits have remained on *DC* since its 2012 launch, and they are not typically edited or updated after publication.

Our study also responds to the "curiously North American and Eurocentric view" (Franklin 2008, 631) of the media that has prevailed in journalism studies (Hanusch 2017, 390). Relatively few studies on digital media in India have been published, with little to no mention of alternative and Dalit news outlets such as *DC*. Despite overwhelming corruption of mainstream Indian media characterized by the "Murdochization" of the nation's journalism, Sonwalkar (2017) notes "an encouraging aspect is the corrective role played by users of digital media to instantly point out errors of fact, bias, and perspective in mainstream journalism," roles ostensibly embodied by *DC* (529). Indeed, *DC* represents the apotheosis of how, since 2015, "several news websites were providing an alternative to the dumbing down of news in the mainstream news media [in India] by focusing on a range of opinions, news, and features." *DC* leverages increasing internet penetration in India to extend "forms of 'recognition' to regions and minorities," in this case Dalits, "that are usually marginalized or ignored in mainstream news discourse" (Sonwalkar 2017, 530).

We are, however, careful not to exaggerate the potential of "Dalit media" because, as Bathran (2016) noted, considering Dalits as a monolith and unfailingly in contradistinction to upper castes fails to address the issue of discrimination within Dalits. An examination of *DC's* blog and YouTube channel shows that Dalits from certain regions of the country such as Northeast India have received little archival space.

Furthermore, because *DC* is currently based out of three metropolitan cities, there does exist an urban–rural divide, although to a much lesser extent than in mainstream media. These exclusions serve a constant reminder against uncritical celebration of Dalit digital media. Lack of participation from a certain portion of Dalits, whether based on region, class, gender or language differentiation, would imply further social exclusion as the internet becomes the dominant domain of public sphere (Sreekumar 2007).

It must also be added that the digital space where *DC* is hosted is controlled by private corporations and the government, which makes it possible to restrict and block access to the archive. In fact, in February 2017, YouTube took down *DC*'s channel for a day after an alleged "copyright violation" (Express News Service 2017) on one of the *DC* videos pertaining to Dalits' right to beef consumption, which has become a topic of intense political debate in contemporary India under the rule of the BJP government whose votaries have publicly lynched Dalits and Muslims over the past few years for eating beef (Nair 2017). Internet shutdowns and censorship on free speech have in general increased in India (Agarwal, Bhatnagar, and Goyal 2018) with the ruling government's tendency to confront problems of multi-community, multi-religious democracy through censorship (Schulz 2016). And, as the government's indefinite blocking of the Dalit portal [dalitstan.org](http://dalitstan.org) illustrates, marginal voices on the web are particularly susceptible to censorship.

The potential for *DC* to influence substantive institutional progress depends on the government's responsiveness and openness to change. According to Castells (2015), "The more the state is responsive to the demands of society, the lesser is the intensity of autonomous social movements." The inverse is also proportionally true, as in the case of *DC*. When the state is less responsive, social movements reach a higher pitch of intensity. The high intensity of *DC* as an autonomous social movement is a direct measure of the Indian government's unresponsiveness and insensitivity to issues of caste, whose chronic and protracted neglect is now exposed in unprecedented detail through the hard glare of the online archive. "When social movements do exist and the state institutions are open to change, the transformative potential of social movements may find an institutional expression" (Castells 2015, 275). The lack of openness to change in India's ruling government, however, and its tightening grip on the media over the past few years suggests institutional change will encounter severe—if not insurmountable—resistance in the immediate future. Despite this major obstacle, the spreadable nature of the archive on social media may apply pressure to reform as the Indian government seeks international credibility in order to rise from a developing nation to a major player in the global economy.

In addition to its function as agent of social and political reform, the digital archive of *DC* has progressive implications for academic historical research and scholarship. In previous historical scholarship, the term "subaltern," used by Antonio Gramsci to describe subordination in terms of class, caste, race, gender, language, and culture, has reified peasant rebellion in India into a monolithic faceless identity. In the absence of their own self-generated historical record, "the peasant rebel" in such cases "has been dealt with merely as an empirical person or member of a class, but not as an entity whose will and reason constituted the praxis called rebellion," as Subaltern Studies scholar Ranajit Guha (1983) aptly observes (2). Whereas scholars have relied upon post-structuralist discourse analysis methods treating official sources and documents'

rhetorical uses of the subjugated classes, the digital archive of *DC* offers a new set of “data” generated by Dalits themselves. For centuries, Indian peasants and Dalits left few documents such as worker’s diaries testifying to their experience in their own voice. *DC* renders the plight of the outcastes as a potent signifier of shared experience. As such, the archive is rich with evidence of online political community formation and digital sociality defining the Dalit not just in terms of their own troubled past and ongoing present, but also as a network of agents responding directly to the repressive forces of the elite.

Through intimate raw footage and documentary video, this grassroots digital archive has the power to transform Subaltern Studies by restoring “the subaltern’s autonomous consciousness” which has been systematically denied in depictions by Western social histories. The humanist-subject agent as represented in the digital archive can transcend the historical narrative of failure associated with transient and finite uprisings, instead pointing to the ongoing act of online rebellion through this distinctive network of outrage and hope, as Castells (2015) would have it, a record of something much more lasting than a “fleeting moment of defiance” (Prakash 1994, 1480). In the process, *DC* has destabilized the notion of the archive itself, specifically by undermining its association with state-sanctioned bureaucratic structure and strictures of archival science. *DC* illuminates the Dalit desire to archive, to chronicle, to “seek some kind of immovable historical foundation” (Yale 2015, 334). The voice of the Dalits has never been more thoroughly documented as on *DC*, a collection of personal archives and chronicles forming a network of social and political advocacy.

### Geolocation of Study

India.

### DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

### REFERENCES

- Agarwal, Sheetal D. and Michael L. Barthel. 2015. “The Friendly Barbarians: Professional Norms and Work Routines of Online Journalists in the United States.” *Journalism: Theory, Practice & Criticism* 16 (3): 376–391.
- Agarwal, Simran, Rahul Bhatnagar, and Shilpi Goyal. 2018. “Internet Shutdowns Become Chronic.” *The Hoot*, January 5. <http://www.thehoot.org/media-watch/digital-media/internet-shutdowns-become-chronic-10463>
- Apparasu, Srinivasa Rao. 2017. “Rohith Vemula not a Dalit, Belonged to Other Backward Class: Andhra Govt.” *Hindustan Times*, January 17.
- Aravind, Indulekha. 2016. “How Online Anti-Caste Platforms are Reclaiming and Reasserting Dalit Space.” *The Economic Times*, August 14.
- Bandyopadhyay, Ritajyoti. 2011a. “Politics of Archiving: Hawkers and Payment Dwellers in Calcutta.” *Dialectical Anthropology* 35 (3): 295–316.

- Bandyopadhyay, Ritajyoti. 2011b. "A Historian Among Anthropologists: Comments on 'Politics of Archiving.'" *Dialectical Anthropology* 35 (3): 331–339.
- Bathran, Ravichandran. 2016. "The Many Omissions of a Concept: Discrimination Amongst Scheduled Castes." *Economic and Political Weekly* 51 (47): 30–34.
- Belair-Gagnon, Valerie, Smeeta Mishra and Colin Agur. 2014. "Reconstructing the Indian Public Sphere: Newswork and Social Media in the Delhi Gang Rape Case." *Journalism: Theory, Practice & Criticism* 15 (8): 1059–1075.
- Bhim, Jai. 2018. "About Us." *Dalit Camera*. Accessed 11 June 2018. <http://www.dalitcamera.com>
- Borger, Merel, Anita Van Hoof, and José Sanders. 2016. "Expecting Reciprocity: Towards a Model of the Participants' Perspective on Participatory Journalism." *New Media & Society* 18 (5): 708–725.
- Bødker, Henrik. 2017. "The Time(s) of News Websites." In *The Routledge Companion to Digital Journalism Studies*, edited by Bob Franklin and Scott Eldridge II, 55–63. New York: Routledge.
- Bødker, Henrik, and Niels Brügger. 2018. "The Shifting Temporalities of Online News: The *Guardian's* Website from 1996 to 2015." *Journalism: Theory, Practice & Criticism* 19 (1): 56–74.
- Bourmans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit: An Overview of Relevant Content Analysis Approaches and Techniques for Digital Journalism Scholars." *Digital Journalism* 4 (1): 8–23.
- Campbell, Vincent. 2015. "Theorizing Citizenship in Citizen Journalism." *Digital Journalism* 3 (5): 704–719.
- Castells, Manuel. 2015. *Networks of Outrage and Hope: Social Movements in the Internet Age*. Cambridge: Polity Press.
- Chadha, Kalyani, and Linda Steiner. 2015. "The Potential and Limitations of Citizen Journalism Initiatives: Chhattisgarh's CGNet Swara." *Journalism Studies* 16 (5): 706–718.
- Chopra, Rohit. 2006. "Global Primordialities: Virtual Identity Politics in Online Hindutva and Online Dalit Discourse." *New Media & Society* 8 (2): 187–206.
- Cook, Terry. 2001. "Archival Science and Postmodernism: New Formulations for Old Concepts." *Archival Science* 1 (3): 3–24.
- Couldry, Nick. 2008. "Digital Storytelling, Media Research and Democracy: Conceptual Choices and Alternative Features." In *Digital Storytelling, Mediatized Stories: Self-Representations in New Media*, edited by In Knut Lundby, 41–60. New York: Peter Lang Publishing.
- Cutrona, Carolyn and Daniel Russell. 1990. "Type of Social Support and Specific Stress: Toward a Theory of Optimal Matching." In *Social Support: An Interactional View*, edited by Barbara Sarason, Irwin Sarason, and Gregory Pearce, 319–366. New York: Wiley.
- Dash, Bidu Bhusan. 2013. "Temple Entry in Odisha by the Dalit: An Ethnographic Study of Media Articulation." *Asia Pacific Media Educator* 23 (1): 63–84.
- Deuze, Mark. 2008. "The Changing Context of News Work: Liquid Journalism for a Monitorial Citizenry." *International Journal of Communication* 2: 848–865.
- Dhillon, Amrit. 2014. "YouTube Eye on a Brutal Caste System." *The Sydney Morning Herald*. April 13. <https://www.smh.com.au/world/youtube-eye-on-a-brutal-caste-system-20140412-36k3u.html>

- Express News Service. 2017. "YouTube Flip Flop: Dalit Camera Taken Down, Then Restored." February 1, *The New Indian Express*. Accessed 17 December 2017. <http://www.newindianexpress.com/states/telangana/2017/feb/01/youtube-flip-flop-dalit-camera-taken-down-then-restored-1565679.html>
- Franklin, Bob. 2008. "The Future of Newspapers." *Journalism Practice* 2 (3): 306–317.
- Gautama Dharmasutra. 1999. "Dharmasutra." In *Dharmasutra: The Law Codes of Ancient India*, translated and edited by Patrick Olivelle, 74–126. Oxford: Oxford University Press.
- Guha, Ranajit. 1983. "The Prose of Counter-Insurgency." In *Subaltern Studies: Writings on South Asian History and Society*, edited by Ranajit Guha, 1–15. Oxford: Oxford University Press.
- Hartley, John. 2012. *Digital Futures for Cultural and Media Studies*. Chichester, WS: Wiley Blackwell.
- Harvey, David. 2010. *The Enigma of Capital*. London: Profile Books.
- Hanitzsch, Thomas, Folker Hanusch, Claudia Mellado, Maria Anikina, Rosa Berganza, Incilay Cangoz, Mihai Coman, et al. 2011. "Mapping Journalism Cultures across Nations: A Comparative Study of 18 Countries." *Journalism Studies* 12 (3): 273–293.
- Hanusch, Folker. 2017. "Transformations of Journalism Culture." In *The Routledge Companion to Digital Journalism Studies*, edited by Bob Franklin and Scott Eldridge II, 389–395. New York: Routledge.
- Jeffrey, Robin and Assa Doron. 2012. "Mobile-izing: Democracy, Organization and India's First 'Mass Mobile Phone' Elections." *The Journal of Asian Studies* 71 (1): 63–80.
- Jeffrey, Robin. 2001. "[Not] Being There: Dalits and India's Newspapers." *South Asia* 24 (2): 225–238.
- Khan, M Ghazali. 2016. "Justice for Rohith Vemula—Killed by Institutional Brahmanical Casteism! Candlelight Vigil in London." South Asia Solidarity Group, January 28.
- Matzner, Deborah. 2014. "*Jai Bhim Comrade* and the Politics of Sound in Urban Indian Visual Culture." *Visual Anthropology Review* 30 (2): 127–138.
- Meadows, Michael. 2013. "Putting the Citizen back into Journalism." *Journalism: Theory, Practice & Criticism* 14 (1): 43–60.
- Mehta, Vanya. 2014. "YouTube Channel Become Rallying Point in India's Dalits." *BBC*, January 7.
- Mishra, Vinod. 1999. "The Dalit Question." In *Selected Works, (n.e.)*, 192–194. New Delhi: Central Office of the CPI (ML).
- Mitra, Ananda. 2001. "Marginal Voices in Cyberspace." *New Media & Society* 3 (1): 29–48.
- Mody, Bella. 2015. "How well do India's Multiple Language Dailies Provide Political Knowledge to Citizens of this Electoral Democracy?" *Journalism Studies* 16 (5): 734–749.
- Nagpaul, Dipti. 2017. "Write Back in Anger." *The Indian Express*, January 31.
- Nair, Supriya. 2017. "The Meanings of India's 'Beef Lynchings'." *The Atlantic*. Accessed 17 December 2017. <https://www.theatlantic.com/international/archive/2017/07/india-modi-beef-lynching-muslim-partition/533739/>
- Pandian, Mathias Samuel Soundra. 2007. *Brahmin Non Brahmin: Genealogies of the Tamil Political Present*. New Delhi: Permanent Black.

- Pandian, Mathias Samuel Soundra. 2008. "Writing Ordinary Lives." *Economic and Political Weekly* 43 (38): 34–40.
- Paul, Subin. 2017. "Between Participation and Autonomy: Understanding Indian Citizen Journalists." *Journalism Practice* 12 (19): 1–17. doi:10.1080/17512786.2017.1331707.
- Poell, Thomas and Sudha Rajagopalan. 2015. "Connecting Activists and Journalists: Twitter Communication in the Aftermath in the 2012 Delhi Rape." *Journalism Studies* 16 (5): 719–733.
- Povinelli, Elizabeth A. 2011. "The Women on the Other Side of the Wall: Archiving the Otherwise in Postcolonial Digital Archives." *A Journal of Feminist Cultural Studies* 22 (1): 146–171.
- Prakash, Gyan. 1994. "Subaltern Studies as Postcolonial Criticism." *The American Historical Review* 99 (5): 1475–1490.
- Prasad, Chandra Bhan. 2017. "But the Earth Moves, And Rohith Vemula is a Dalit." *The Indian Express*, August 17.
- PTI. 2016. "Vemula's Suicide an 'Institutional Murder': Satchidanandan." *The Hindu*, January 23.
- Rao, Shakuntala and Vipul Mudgal. 2015. "Introduction." *Journalism Studies* 16 (5): 615–623.
- Rao, Shakuntala and Herman Wasserman. 2015. "A Media Not for All." *Journalism Studies* 16 (5): 651–662.
- Schulz, Suzanne. 2016. "Temporary Bans and Bad Laws: The Aarakshan Ban and the Logics of Censorship in India." *Communication, Culture & Critique* 9 (4): 537–557.
- Shirky, Clay. 2009. *Here Comes Everybody: How Change Happens when People Come Together*. New York: Penguin.
- Shobana, Nidhin. 2016. "Rohith Vemula Tribute." *Dalit Camera*. Accessed 31 December 2017. <http://www.dalitcamera.com/rohith-vemula-tribute/>
- Sonwalkar, Prasun. 2017. "A Conundrum of Contrasts: The 'Murdochization' of Indian Journalism in a Digital Age." In *The Routledge Companion to Digital Journalism Studies*, edited by Bob Franklin and Scott Eldridge II, 528–536. New York: Routledge.
- Spivak, Gayatri. 1985. "The Rani of Sirmur: An Essay in Reading the Archives." *History and Theory* 24 (3): 242–272.
- Sreekumar, T. T. 2007. "Cyber Kiosks and Dilemmas of Social Inclusion in Rural India." *Media, Culture & Society* 29 (6): 869–889.
- Stoler, Ann Laura. 2002. "Colonial Archives and the Arts of Governance." *Archival Science* 2: 87–109.
- Tharakam, Bojja. 2017. Bojja Tharakam Speaks on Rohith Vemula. *Dalit Camera*. Accessed 31 December 2017. <http://www.dalitcamera.com/bojja-tharakam-speaks-on-rohith-vemula/>
- Thirumal, P. 2008. "Situating the New Media: Reformulating the Dalit Question." *South Asian Technospaces* 36: 98–122.
- Thirumal, P. and Gary Michael Tartakov. 2011. "India's Dalits Search for a Democratic Opening in the Digital Divide." In *International Exploration of Technology Equity and the Digital Divide: Critical, Historical and Social Perspectives*, edited by Patricia Randolph Leigh, 20–39. Hershey, NY: Information Science Reference (IGI Global).
- Thompson, Edward. 1966. "History from Below." *Times Literary Supplement* 65: 279–280.
- Udapa, Sahana. 2016. "Archiving as History-Making: Religious Politics of Social Media in India." *Communication, Culture & Critique* 9 (2): 212–230.



- Vemula, Radhika. 2017. "Radhika Vemula's Speech at the 10th DYFI All India Conference, Kochi, Kerala." *Dalit Camera*. Accessed 31 December 2017. <http://www.dalitcamera.com/radhika-vemulas-speech-10th-dyfi-india-conference-kochi-kerala/>
- Wall, Melissa. 2015. "Citizen Journalism." *Digital Journalism* 3 (6): 797–813.
- Weld, Kirsten. 2014. *Paper Cadavers: The Archives of Dictatorship in Guatemala*. Durham, NC: Duke University Press.
- Widholm, Andreas. 2016. "Tracing Online News in Motion: Time and Duration in the Study of Liquid Journalism." *Digital Journalism* 4 (1): 24–40.
- Yale, Elizabeth. 2015. "The History of Archives: The State of the Discipline." *Book History* 18 (1): 332–359.

# DIGITAL ARCHIVES AS SUBALTERN COUNTER-HISTORIES

## Situating “Favela Tem Memoria” in the Rio de Janeiro media and political landscape

**Stuart Davis**

*This piece analyzes the possibilities raised by leveraging digital archives as tools for community advocacy in marginalized communities. Drawing upon a case study of “Favela Tem Memoria”, an archive created by the Viva Rio NGO, I will first analyze how the project offers a digital space for collecting and sharing a “counter-history” of these communities that have been historically marginalized in economic, social, and political terms. Then, I will discuss how the collapse of the archive in 2017 reflects a larger problem within NGO-launched and maintained projects regarding the tension between meeting donor needs and community needs.*

### **Introduction: The History of “Favela Tem Memoria” as Archive and Intervention**

In 2005, Viva Favela, the media production branch of Viva Rio, the oldest non-governmental organization focused on digital media production training in the favelas (or unincorporated urban peripheries) of Rio de Janeiro, launched “Favela tem Memoria”, an online archive housing pictures, photographs, and oral histories of community residents. Beyond this preservationist element, the project also trained residents to produce their own stories as a way to create cultural and political linkages between the past and present (Viva Rio 2009). Featuring four major sections labeled “Favelas” [“Neighborhoods”], “Depoimentos” [“Testimonies”], “Favelese” [“Favela Lingo”], and “Fotos” [“Photographs”], the website housed materials produced by participants from these geographically, economically, and politically marginalized communities. To create the archive, Viva Favela staff members reached out to community members within 10 favela neighborhoods through online advertisements on the NGO’s site and presentations at community centers. After a training process, participants were given access to the archive for uploading materials. The site contained four types of materials: photographs uploaded by community members or donated to the project and

uploaded by staffers, oral histories recorded and transcribed by staffers, written narratives produced and uploaded by individuals, and special sections created by staff (namely the “Favelese” lingo section). The basic idea behind the project was to collect as much history about the neighborhoods as possible to act as a resource for community members, the public, and journalists interested in using the digital archive as a tool for materials to help create richer and more nuanced narratives about favelas (Agência Brasil 2016).

For roughly 10 years, “Favela Tem Memória” acted as a repository for creating and spreading histories of favela communities. In 2015, activists used videos produced by “Favela Tem Memória” as part of a successful anti-eviction land titling project in the Salgueiro favela. In November 2016, the project partnered with MetrôRio, the public–private partnership corporation that runs the city’s subway system, and the think-tank Instituto Invepar to create a traveling exhibit showcasing 228 photographs with explanatory didactic panels from the “Favela Tem Memória” site (Agência Brasil, 2016). Traveling between train stations located in neighborhoods close to favelas, the idea behind the exhibition was to “show commuters and others using the Metro that favelas had as much of a history in the city as any other community” (Viva Rio, 2016). Accompanied by media coverage and multiple press releases from Viva Favela, the exhibit traveled across Metro stations citywide from February–March, 2016.

From the perspective of what the archive accomplished, I argue that “Favela Tem Memória” provides a powerful example of what Stuart Hall calls a “living archive” (Hall 2000). The project worked with local residents’ associations within the favelas in two distinct but related activities: it first helps them curate materials into online exhibitions commemorating significant cultural and political events in their neighborhoods while at the same time allowing them to include their own personal voices through. From this perspective, I argue that this digital archive serves the double function of memorializing histories of favela life almost completely obscured in the larger history of Rio de Janeiro while promoting advocacy on behalf of Rio’s favela residents in both the long term (by challenging stereotypes promulgated by mainstream media and short term (through providing resources used in legal and political struggles). Through these activities the archive acts as both a space for preserving and celebrating subaltern histories and a tool for promoting social change within contemporary Rio de Janeiro.

Addressing the relationship between the materials disseminated by “Favela Tem Memória” and the mobilization of the site for advocacy practices provides two conceptual interventions with potential political ramifications. Within scholarship on digital humanities and media studies (specifically scholarship related to digital media as tools for preserving “popular” history), the project raises key geopolitical points about the way that online archives can be used to engage in the act of *counter-historiography* or the re-writing of the historical archive from the perspective of subaltern or oppressed perspectives. Crucially, this process of re-writing serves the double purpose of preserving community history for future generations while also potentially offering a tool for activists interested in leveraging the material as evidence that can be used to advocate for favelas. This double function of the archive in “Favela Tem Memória” serves to both capture the histories of lives covered up in the “official” histories of favelas as a resource for producing counter-power while at the same time leveraging interactive participation to empower users as media makers, favela residents as user of the site/

audience members of the presentation, and communities as beneficiaries from the use of materials archived in the project in court battles over property rights.

While the majority of this discussion will focus on the avenues provided by the archive/exhibit for the lived histories of favela residents to counter dominant media narratives and bolster political and legal struggles, the conclusion will briefly address the aftermath of the 2016 Metro exhibits and subsequent media coverage. During the course of the 2016 exhibit, the “Favela Tem Memoria” stopped displaying original content by residents. In October 2017, the site went offline without any warning or fanfare (only displaying a domain name system (DNS) error). At the time of writing this article, no one from the NGO could be reached for technical support. This turn in the history of the archive drives the conclusion of this piece, an analysis of how the political economy of NGOs creates a lack of accountability over management and preservation of archives that leaves participants without control of contributing materials or accessing materials previously posted.

### **Representations of Favelas within Brazilian Media: A History of Spectacularized Violence and Criminality**

In order to understand why the project initiators of the “Favela Tem Memoria” archive launched the project, it is important to briefly address why this intervention into the representation of these spaces was then and continues to be necessary. Since the rise of drug trafficking and the amplification of the small arms trade within favelas in the late 1970s, Brazilian mainstream news media have perpetually represented these areas through the related tropes of violence and abject poverty. Sensationalistic articles about gunfights, hijacking, and drug abuse appear on a regular basis as well as bombastic television news programs like *Cidade Alerta* [*City on Alert*] that often feature real-time shootouts and police invasions into favelas (e.g. Mayer 2006; Matheus 2011). One of the central tropes characterizing media portrayals of favelas is the myth of “balas perdidas” [“lost bullets”]. This narrative, popularized in the 1990s by news media in Rio, claims that armed violence between different trafficker factions and between traffickers and the police is so constant that there was a persistent danger of stray bullets incurring collateral damage (Agência Estado 2001). Consequently, the security option often proposed during the 1980s–early 2000s was to invade or forcibly pacify favela communities without regard to the safety of occupants not involved in drug trafficking or paramilitary activity (Leite 2012).

In response to the humanitarian crises generated by the wanton violence and discrimination practiced against favela residents, the 1990s witnessed the growth of a large number of NGOs aimed at using video production, music, photography, and (eventually) web-based media production as tools to create a counter-narrative against those offered by mainstream media outlets and municipal police (Platt and Neate 2006; Davis 2016). Most of the projects that came out of this flourishing of NGO-facilitated training programs focused on training favela residents (usually younger residents) to create media documenting their daily lives, cultural activities unique to favelas, and other elements related to the non-violent and experientially rich nature of these communities. These materials would then be circulated via photo exhibits, film festivals, or online spaces. However, Viva Favela’s “Favela Tem Memoria” project aimed at

explicitly at recruiting middle-aged or older members of the community to contribute oral histories, photographs, or other pre-existing materials to an online archive. As an archive designed around collecting and curating residents' memories of their communities, this project combines the experiential element of other favela-based media projects with an archival impulse aimed at preserving the past as a way to valorize the present.

### **The Digital Archive as Subaltern Counter-History and Lived Cultural Resource**

The innovative aspect of "Favela Tem Memoria" as a repository for historical materials and a tool for advocacy comes through its ability to create juxtapositions between documents acting as snapshots from multiple time periods. However, the medium specific elements of digital archives differentiate them from more traditional paper archives. When attempting to understand the potential for undermining generations of historical stereotypes and deleterious representations, "Favela Tem Memoria" must be situated within a certain form of historical work. Moving outside the scholarly parameters of media studies and digital humanities, critical literature within the sub-field of historiography addresses precisely the geopolitical dynamics present in the creation of historical archives. In particular, the configuration of scholars working in Southern Asia and Hispanophone Latin America under the collective characterization of "subaltern studies" provide two related concepts for framing the role of archives in advocacy projects working with marginalized populations: decolonial counter-history and "living" counter-history. The decolonial approach emphasizes the political import of including knowledge from historically marginalized populations while "lived archeology" is linked to the ways that contemporary experience can serve as a springboard for reflection on the historical past. I argue that a combination of these two elements provide the theoretical basis for understanding the counter-history created and mobilized in "Favela Tem Memoria".

The term "decolonial thinking" is generally associated with the work of anthropologist, linguist, and cultural historian Walter D. Mignolo. His work attempts to promote knowledge systems (including customs, languages, and artistic practices) occluded by the global rise of Western Modernity. Arguing against the way that both capitalist and Marxian orthodoxies presuppose a total fissure between a "pre-modern" period governed by localism coupled with technological backwardness and a "techno-modernity" characterized by highly developed political, economic, and material conditions (e.g. Meiskins Wood 1991), Mignolo and others working within this theoretical-historical framework continually push for the interrogation of "non-Modern" commonly accepted narratives. Beginning with *Global Elements of Literary Theory* (1978), his corpus is built around connecting knowledge production and geopolitical power (Mignolo 2012). For Mignolo, cultures of scholarship "become part of a political domain of discourses and social concerns, coupled with knowledge oriented toward emancipation/liberation" (Mignolo 2012, 107). Within this political discourse, the work of decolonial thinking consists of locating situations where non-Western histories produce a "different kind of rationality" than that famously espoused by Max Weber in his now iconic work on the development of occidental institutional culture (Mignolo 2012, 187; 188–191 *passim*).

Mignolo argues that including histories of subaltern populations is critical for breaking down dominant narratives of historical progress that situate marginalized identities as either antecedents to contemporary society (in the case of indigenous communities) or as an underclass who do not fit within the current global zeitgeist (in the case of ethnic or racial minorities, geographically marginalized groups, or the urban poor).

As a practice of re-writing history this decolonial approach serves the crucial political function of tracking why certain subaltern narratives are covered at certain times. Hence one of the main goals of this approach is to analyze how “renaming” history serves as a part of an attempt to re-write the past in a way that is more palatable to dominant forces within society. For example, historian Shahid Amin’s *Event, Metaphor, Memory: Chauri Chaura 1922–1992* (1995) examines how a peasant protest in rural South Asia originally launched as part of Mahatma Gandhi’s campaign against British home rule gets represented differently throughout the 20th century as it is initially praised as a sign of popular disapproval for the British then condemned as “mob violence” when Gandhi’s movement attempts to legitimize itself as anti-violence and finally celebrated in an overtly nationalist way as a moment in the blossoming of the “modern” nation of India. Importantly, proponents of subaltern studies do not represent just an example of uncovering secret histories but re-writing existing histories of Rio’s favelas. Arjun Appadurai’s “Archive and Aspiration” explicitly connects digital archives to this process of counter-history as he argues that “newer forms of electronic archiving restore the deep link of the archive to popular memory and its practices, returning to the non-official actor the capability to choose the way in which traces and documents shall be formed into archives ...” (2003, 18).

At the same time as it was trying to produce a counter-history of the favelas, the project was also invested in recording, preserving, and disseminating cultural and political activities in the *present*. Returning to Hall’s notion of the archive as a syncretic attempt to connect past practices with current political, economic, or health problems facing neighborhoods, the goal of a digital archive is to both encapsulate the past and to create connections with the present. Writing in the context of diasporic communities, Hall posits a definition of “living” grounded in multi-temporality: “‘Living’ means present, on-going, continuing, unfinished, open-ended ... This notion of ‘living’ is strongly counter-posed to the common meaning accorded to ‘tradition’, which is seen to function as the prison-house of the past” (Hall 2001, 89). The way he conceptualizes the “living” as a porous temporality where past and present intersect or overlap provides a useful way for thinking about “Favela Tem Memoria” as grounded both in community history and community present. Through combining photographs from different decades between the 1950s and the 2000s along with narratives from community members of various ages, the archive manifests a sense of temporal co-existence between different periods of the favelas.

The way that “Favela Tem Memoria” captures this sense of “temporal co-existence” offers its most unique contribution to scholarly and practitioner literature. Though similar to other digital journalism sites that are structured around user-generated content, the project also provides new avenues for marginalized or subaltern populations to present not just their own individual perspectives but also those of their family members and community. In this way the archival nature of the project adds another facet to the process through which digital technologies expand the field of journalism to include the



**FIGURE 1**

Screen capture of the homepage for “Favela Tem Memória”. The sections of the archive run across the top of the page. From left to right, the sections read “Sobre Nós” [“About Us”], “Favelas”, “Depoimentos” [“Testimonials”], “Favelês” [“Favela Slang”], “Fotos” [“Photographs”], “Parceiros” [“Partners”], and “Contato” [“Contact”]

narratives of everyday citizens (Thurman 2008; Mpofu 2014). “Favela Tem Memória” presents the case that digital technologies do not just allow for the recording and dissemination of lived experiences of individuals but also memories of the social world they inhabit.

### Research Approach

Though the crashing of the site that housed “Favela Tem Memória” prohibited an in-depth analysis of the site in extensive detail, the manner in which the site functions as a resource for empowerment and advocacy can be detailed through a qualitative textual analysis of materials available on the site before it disappeared, videos produced in 2015 as part of its anti-eviction campaign (which are still available on the Viva Favela YouTube channel), and the 17 articles published about the 2016 Metro exhibit. Analyzing these three elements in conversation together begins to elucidate how “Favela Tem Memória” might be conceptualize as simultaneously a resource for capturing, preserving, and proliferating favela history and contemporary lived experience. Though the closure of the digital archive’s site occludes a more comprehensive analysis, the following sections will address both key elements of the archive that fostered user involvement and the impact of the archive on both media coverage of favelas and on specific legal struggles within favelas. Aimed at interrogating how the project attempts to empower users through producing and curating personal histories, the first section will discuss the planning, design, and specific sections of the forward-facing “Favela Tem Memória” site (Figure 1 offers an image of the homepage highlighting the archive’s different categories). The bulk of the analysis in this section will address in detail how the “Photographs” and “Testimonial” categories of the site are designed to create a sense of chronological continuity between the 1960s and the 1970s and present in order to offer a history of community solidarity and struggle within favelas. The second section of the discussion will look at the impact the archive and 2016 exhibit

instigated beyond the confines of cyberspace. More explicitly, this section will address media coverage of the exhibit as well as the role of materials from the archive used in a court case to determine legal occupancy for a group of favela residents. After discussing the site's role in creating new histories of favelas and mobilizing those histories for advocacy projects, the discussion will conclude by balancing these progressive changes on the behalf of favelas with the NGO's apparent decision to stop updating and eventually close the archive.

### **“Favela Tem Memoria” as Living Archive: Creating Subaltern Counter-Histories**

“Favela Tem Memoria” was launched in 2005 by Viva Favela, the media production wing of the larger non-governmental organization (NGO) Viva Rio. Launched in 2001, Viva Favela's goal was to use media production to provide audio-visual avenues for celebrating and promoting favela culture in order to create a new historical narrative about these areas that was explicitly oppositional to mainstream narratives painting them in highly racialized and criminalized fashion. It is crucial to note that the project was initially set up as a reaction to negative media representations. In this spirit, the project began as a channel for critiquing the way “professional” journalists working for Rio-based and larger newspapers, television programs, etc. depicted favelas as spaces of criminality and violence by focusing almost exclusively on military confrontations between the police and the drug traffickers exercising political control over the neighborhood (Jovchelovitch and Priego-Hernández 2013; Baroni 2013; Guedes Rocha 2016). Started by Viva Favela staff in 2015, “Favela Tem Memoria” built upon this commitment to using digital media as a tool for both preserving and potentially amplifying the perspectives of the average residents of these communities in Rio.

In line with Mignolo's assertion that producing counter-narratives constitutes a political act and Hall's claims that subaltern archives are better understood as living multi-temporal organisms instead of static collections, it is crucial to address how the “Favela Tem Memoria” differs from what one would consider traditional archives. Creative industries scholar Jean Burgess's (2007) notion of “vernacular creativity” provides a useful concept for understanding what this archive is attempting to capture: “Vernacular creativity refers to the variety of everyday creative practices like storytelling, family photographing, scrapbooking, journaling and so on that pre-exist the digital age and yet are co-evolving with digital technologies and networks in really interesting ways” (Burgess 2007, 14). In examining the history of this concept, Burgess argues that digital media have greatly amplified the ability of local communities to preserve and disseminate these traditions without having to turn to interlocutors. The tradition of vernacular creativity via digital media production has a history in favelas that goes back almost 20 years. Since the creation of the earliest Favela-based blogs sites in the early 2000s, residents have turned to first participatory websites and then social media as a channel to reach larger municipal, national, and global environments. In the 2000s, participatory sites like the one operated by Viva Favela provided platforms for favela residents to upload videos, photographs, or written narratives. Along with these NGO-run sites, neighborhood-specific blogs, Facebook pages, and Twitter accounts





**FIGURE 2**

Screen capture of photograph from “Favela Tem Memória” photo bank showing the eviction of Chapeu Mangeuira Favela, circa 1950s. Courtesy of Viva Favela

have expanded at an incredible rate over the last decade. The largest favela-specific Twitter profile, “Voz das Comunidades” [“Voice of the Favelas”], now has over 350,000 followers. As a fundamentally participatory site, “Favela Tem Memória” is clearly part of the same trend as these other projects. However, as an archive it also adds an attention to historical preservation and curation not necessarily present in other types of digital media projects focusing on favelas. As a hybrid archive/participatory media site, the project inserts a powerful democratizing element to the production of history through archival documentation.

The largest element of the “Favela Tem Memória” archive was the photographic archive curated on the “Fotos” section of the site. To create this section, Viva Favela team members solicited donations from 10 different communities from different parts of the municipality of Rio and in surrounding areas. When asking for photo donations, NGO workers explicitly targeted photographs from the 1960s, 1970s, and 2000s/2010s. The archive contained no photos before 1960 or from the 1980s or 1990s. The decision to target these two time periods was neither capricious nor careless. The 1960s–1970s represented the period when favelas began to spring up across the nation. As industrialization intensified nationwide, migrants from the Northeast part of the nation flooded into Rio, São Paulo, and similar metropolitan areas for better job opportunities (Perlmann 1980). In the photo archive, this period is thematized by two different types of pictures: those depicting community building and those depicting community responses to crises. Organized by neighborhood (e.g. Borel; Favela da Ramos), the first type of photos focuses on the construction of houses, typical daily scenes like women hanging laundry and children playing in the street, and general scenes of shared public space. The second type of photo display specific crisis situations faced by community residents. Organized by crisis events including flooding and forced eviction, photos in this group display one of two conceptual themes: “collaboration” or “dejection”. The photos depicting “collaboration” show residents working together in activities such as

helping each other escape from natural disasters or defend themselves from police attack. The “dejection” photos offer powerful encapsulations of what documentary scholar Paula Rabinowitz has called “the social totality in a single image” (Rabinowitz 1994, 38). This group of photos presents residents looking forlorn as police officers, bulldozers, and construction crews clear out their communities. Figure 2 provides a typical example of this sort of photo: a group of children and a young woman holding a small child walk cautiously through the rubble of their community after their homes had been destroyed. This type of picture attempts to frame the larger phenomenon of eviction as a moral issue by focusing on the human costs of forced removals. The photographs from the 2000s are significantly fewer and organized by neighborhood. Though much less thematically distinct than the 1960s/1970s photos, most of the images depict cultural activities including dances, kite flying, the construction of floats for carnival, and bicycle riding. By focusing on “a vida cotidiana” [“everyday life”], a phrase often invoked by favela activists to combat representations of these areas as dens of criminality and ubiquitous violence (Chagas 2009, 244, en. 2), this type of image promotes the notion that there is less distinction between community life and community values than the stereotypes about values would have the public believe. Entirely absent from this group of photos are any pictures of armed drug traffickers.

It is notable that photographs from the 1980s and 1990s are explicitly absent from the “Favela Tem Memória” archive. Though never mentioned explicitly on the site or any related press materials, it is a reasonable certainty that the Viva Favela chose not to include this period because during this time favelas were experiencing the highest levels of violent confrontation between drug traffickers and State security forces. As Soares (2006), Fischer (2011), and other historians have detailed at length, the early 1980s witnessed a rapid boom in violence as a combination of a boom in the domestic cocaine trade and the fortuitous intermingling of local gang leaders and leftist militant guerrillas in Cândido Mendes and other local prisons. The intersection of these two phenomena led to a rapid and deadly militarization of the favelas. The fact that “Favela Tem Memória” contained no images of this element of daily life resonates strongly with Viva Favela’s mission to create a sense of community history that celebrates their cultural richness and promotes a sense of solidarity among community members (César Fernandes 2010). Acknowledging the difficulties and trauma of this period potentially undermines the site’s role as an advocacy resource for favela communities.

The other central element of the archive was the “Testimonials”, short narratives written by community members with accompanying photographs. This portion of the archive leverages the power of the personal in a way resonant with testimonial literature produced within Latin America in the 1980s–1990s (e.g. Beverly 2004). Among the 28 testimonies, 27 showcased stories about individual residents while one featured an interview with favela historian Rafael Souza. In terms of age distribution, of the 27 participants 10 were older than 80, 7 between 70 and 80, 5 between 60 and 70, 3 between 50 and 60, and 2 under 50. The two most frequently occurring themes in the stories were migration and the preservation and continuation of cultural traditions including religion, music, dance, and food. For example, a 2007 testimony describes the life of “Dona Penha”, an 83-year-old woman who had moved to the Complexo de Maré favela in the 1960s. Maré, a community where all three of the largest trafficking factions have a presence, is often considered to be the most violent favela in Rio (e.g. UOL 2018).

Instead of discussing the violence in her community, her narrative is based on the ways she has preserved and changed her heritage as a Bahian native. Bahia, considered to be the region of Brazil most heavily influenced by African culture, is one of the most culturally distinct regions of the country (Pinho 2004). Other stories focused directly on memories of community life in the mid-20th century with the omission of the 1980s–1990s. One of the testimonials highlighted in the 2016 Metro exhibit illustrates this sense of a shared collective past: “Whenever it was needed there was a neighbor to help with work, renovation, building a house. And we shared almost everything. If someone needed a clove of garlic, we would give them two or three. If we needed rice someone would give it to us. It was a constant exchange of favors. Everyone was poor, life in the favela was difficult, but everyone helped each other” (Quoted in Mackay 2016). This narrative crystalizes the archive’s overall strategy of presenting individual narratives in a way that could be used to create both a sense of historical community solidarity and as a way to delegitimize media and popular narratives about favelas as nothing more than conflict zones or hotbeds for criminal activity.

### **The Impact of “Favela Tem Memoria” as Advocacy Tool: The Continuing Attempt to Re-program Public Conversations About Favelas**

“Favela Tem Memoria” presents an alternative history or retelling of Rio’s favelas constructed around self-sufficiency, collaboration, and social justice. Outside its Internet home, the archive also performs an advocacy function by leveraging the lived experience of individuals in an attempt to reinforce or shed further light on the necessity to promote intervention in a certain area (Tactical Technology Collective 2009, 6). Between 2015 and 2016, the archive was utilized by two initiatives as a tool for building awareness about favela history, and the historical rights of favela citizens. The first, a 2016 exhibit hosted at subway stations across the city, attempted to re-program popular understandings and journalistic coverage of these areas. The second and more directly advocacy initiative was the incorporation of materials from the archive in a 2015 tenant’s rights on behalf of a favela-based samba school. Both of these initiatives provide instances where the materials in the archive become part of larger attempts at promoting cultural awareness of and justice for Rio’s favelas.

Created 11 years after the digital archive launched, the exhibit in the city’s Metro stations was the first time the project received any public or press attention. While Viva Favela had in the past held openings where they displayed photographic and audiovisual materials (including a 2009 nationwide tour where staff members held workshops to accompany the materials exhibited (Lucas 2013)), the 2016 exhibit provided the first real public exposure of the “Favela Tem Memoria” archive. Working with the city of Rio de Janeiro’s municipal works program and the Invepar Institute (a large non-profit focusing on a variety of community development projects), staff members from Viva Favela chose photographs and chunks of testimony from the archive to blow up and create a museum-style exhibit combined with an interactive display that allowed spectators to browse the “Favela Tem Memoria” site. Viva Favela chose to have the exhibit move through different subway stations that were located in close proximity to large favelas. Beginning at the Pavuna station in the city’s north zone, the project traveled until it reached the General Osorio station in the south zone neighborhood of

Ipanema. Over the course of 2 months (between February and March, 2016) subway passengers had the opportunity to use the project as a way to possibly experience some of the sense of community history and solidarity experienced by the actual participants who helped create the digital archive.

The group did not seem to keep any metrics as to how many passersby stopped at or engaged with the exhibit as it traveled through the Metro stations. However, drawing upon media coverage within Rio of the event provides an approximation of the exhibit's larger impact on audiences. As the exhibit moved from station to station, it began to garner considerable press attention. Many of the major news outlets in the city covered the Pavuna opening. A press release circulated by *Veja*, the largest cultural magazine in Rio, described the exhibit as "providing for our social memory a timeline for understanding the ways that the past and the present are intertwined" (Agência Brasil 2016). *Jornal Extra*, the largest circulating newspaper in Rio, dedicated a section to the opening in Ipanema (*Globo Extra* 2016) The Rio Olympics Neighborhood Watch (RioOnWatch) blog, a dual English-Portuguese publication, provided an interpretation of the exhibit that emphasized the way the exhibit wove together a narrative of solidarity that connected the past, present, and future of favelas (Mackey 2016). These are three of the 15 news stories dedicated to the exhibit. I argue that all of these stories provide evidence that the project is achieving goals on behalf of the community. The amount of media coverage potentially indicates a genuine interest by journalists to understand the situation in favelas in a more nuanced and less sensationalist or derogatory manner.

In 2015, one of the testimonies from "Favela Tem Memória" sparked a small campaign that ended with a community receiving ownership of the Samba school in Salgueiro that had been seized by the Rio city government and was slated to be demolished. This case provides a more explicit example of the real-world advocacy potential of the archive than the Metro exhibit. The process started when a group of activists associated with the Brazilian Bar Association started an investigation into who the history of the title of the property (Monteiro 2016). They found that under legislation enacted during the administration of Lula da Silva the neighborhood association met the requirements for titling. Using this information, the neighborhood association was able to claim title and maintain control over the space. This case provides perhaps the clearest example of how "Favela Tem Memória" moved beyond the space of a digital archive and into that of an instrument for community advocacy.

The 2015 Samba School case points directly to the potential of "Favelas Tem Memória" as a tool for promoting advocacy for favela communities. Incorporating material produced in this context provides a form of documentation that invokes the veracity of the experience at hand. User-generated content provides a perspective more closely linked to individual's lived experience than other forms of news media. Hence, products of this type of media production might reinforce a sense of verisimilitude or mimetic relationship between representation and experience that operates according to what historian Joan Scott has famously labeled the "evidence of experience" logic based on literal transparency (Scott 1991, 775). According to this logic, the lived experience of each individual can serve to stand in for the combined experience of a certain group or constituency in a forceful manner. Put more concretely, incorporating personal experience through digital media might greatly amplify and strengthen

the strategic potential of the advocacy message. In the case of "Favela Tem Memoria", the archive worked to both capture a sense of collective solidarity and providing a tool for community advocacy.

### **Conclusion: Archive as Counter-History and Development Initiative**

The 2016 Metro tour represented a high point in exposure for "Favela Tem Memoria" as the archive seemed to be reaching larger and larger audiences through the traveling exhibit. However, during this same period and after the "Favela Tem Memoria" did not record any new entries. Finally, around Fall 2017 the entire online archive contained at [www.favelatememoria.com.br](http://www.favelatememoria.com.br) disappeared, leaving only the message, "Domain Name System (DNS) is Not Functioning Properly". Though some materials from the archive remained on other sites, the majority of materials collected on or created for the site disappeared. Attempted contact with staff members did not merit response. Instead of either celebrating its achievements in terms of amount of material archived or writing it off as a failure, I propose that "Favela Tem Memoria" simultaneously highlights the cultural and political possibilities of creating a space for collecting and sharing subaltern perspectives while also providing a strong case for critically analyzing questions of stakeholder accountability in NGO-initiated projects.

While it would be both cynical and spurious to claim that NGO staff lost interest in maintaining the archive after the high-profile Metro station exhibit in 2016, the success of the mobile display could point to a central critique launched against NGO-led projects in media production (including those creating archives): groups need to show board members, funders, and other members of their network that they are being "effective". As Rodriguez (2007) and other critics have argued (e.g. Elyachar 2011), in many cases non-profit organizations and NGOs focus so heavily on producing "deliverables" or concrete examples that their work is being effective that they lose sight of how the program or project is affecting participants. While the fact that the archive was launched years before the 2016 Metro exhibit illustrates that NGO workers that produced "Favela Tem Memoria" did not intend the project to merely advertise their work, the pressure to prioritize this type of activity could lead to a situation where maintaining the digital archive is no longer the central emphasis of their work. In short, maintenance of the archive seems to not prove a priority for the NGO or funders.

*In Favela: Four Decades of Living on the Edge in Rio de Janeiro* (2010) sociologist, Janice Perlman (the first American academic to conduct research in favelas) appraises the first five years of "Favela Tem Memoria" in a highly positive manner. In a section entitled "Multiple Knowledges and Sense of the Self", she argues that the digital archive is one of a few projects working in favelas that could potentially produce radical change for residents: "If lines in the sand are erased and favela [residents] are able to find a sense of self-worth and self-respect, it will be that much more difficult for those who wish to keep them in a subservient position" (Perlmann 2010, 353). Her assessment of the project frames it as a path for empowering favela residents to believe they have the ability to promote change and fight injustice in their lives. For a population that has been so historically marginalized, this could be revolutionary. Returning to the idea of a subaltern counter-history, we can see how the project provides the foundations for building social

and political consciousness. In this role “Favela Tem Memória” echoes the elliptical yet strong words of Appadurai: “Rather than being the tomb of the trace, the archive is more frequently the product of the anticipation of collective memory” (Appadurai 2003, 24).

With this positive assessment in mind, the site still crashed in 2017. If it is to be a tool of promoting a counter-history, it has to exist for individuals to view and for more favela residents to add their stories and lived experiences. However, in order to continue the project needs financial support. Sociologist Clifford Bob (2006) has persuasively argued that the only difference between a successful and an unsuccessful NGO is the group’s ability to innovatively package its work in a way that attracts and keeps donors. Adapting Bob’s work to the specific context of digital media, Thrall, Stecula, and Sweet (2014) contend that the increasing usage of digital communication strategies by NGOs globally to attract support for their cause has intensified competition among NGOs to produce “deliverables” that might be appealing or attractive to donors. With this economic situation in mind, creating “deliverables” out of the “Favela Tem Memória” archives is completely justifiable. Framed as a part of a community development initiative, the archive must be able to produce materials that the NGO can use to show funders that it is accomplishing the task it is meant to accomplish. In the case of “archive-based development” this means publicity. The tension between the archive’s role as a community resource and its role as a tool for generating financial support for the NGO raises perhaps the most important yet difficult question: To whom does the archive belong? To the participants who create the materials or to the group that is managing the archive? In this case, there appears to be a potential danger that the answer is *not* both.

## REFERENCES

- Agência Brasil. 2016. “Exposição no Metrô Registra o Passado e o Presente das Favelas Cariocas”. Agência Brasil Notícias, February 29, 2016.
- Agência Estado. 2001. “Balas Perdidas Matam Três Moradores de Favela”. O Estado de São Paulo, October 21, 2001.
- Amin, Shahid. 1995. *Event, Metaphor, Memory: Chari Chaura, 1922–1992*. Berkeley, CA: University of California Press.
- Appadurai, Arjun. 2003. “Archive and Aspiration.” In *Information is Alive*, edited by Joke Brouwer and Arjen Mulder, 14–25. Rotterdam: V2\_Publishing/NAI Publishers.
- Baroni, Alice. 2013. “In-Side-Out: Photojournalists from Community and Mainstream Media Organizations in Brazil’s Favelas.” PhD diss., Department of Creative Industries, Queensland University of Technology.
- Beverly, John. 2004. *Testimonio: On the Politics of Truth*. Minneapolis, MN: University of Minnesota Press.
- Burgess, Jean. 2007. “Vernacular Creativity and New Media.” PhD diss., Creative Industries Faculty, Queensland University of Technology.
- Bob, Clifford. 2006. *The Marketing of Rebellion: Insurgents, Media, and International Activism*. London: Cambridge University Press.
- Ceşar Fernandes, Rubem. 2010. “Nossas marcas.” In *Viva Favela*, edited by Mayra Juça, 94–113. Rio de Janeiro: Viva Rio/Editora Olhares.

- Chagas, Viktor. 2009. *Por que é Cidadão o Jornalista Cidadão? História das Mídias e Jornalismo Cidadão de Base Comunitária no Complexo da Maré*. Rio de Janeiro: Fundação Getulio Vargas.
- Davis, Stuart. 2016. "Relocating Development Communication: Social Entrepreneurship, International Networking, and South-South Cooperation in the Viva Rio NGO." *International Journal of Communication*, 10: 42–59.
- Elyachar, Julia. 2011. *Markets of Dispossession: NGOs, Economic Development, and the State in Cairo*. Durham, NC: Duke University Press.
- Fischer, Brodewyn. 2011. *A Poverty of Rights: Citizenship and Inequality in Twentieth Century Rio de Janeiro*. Palo Alto, CA: Stanford University Press.
- Guedes Rocha, Daniela. 2016. "Da Batalha à Guerra do Rio: Uma Abordagem Espaço-temporal da Representação das favelas na Imprensa Carioca." VII Encontro Nacional de Estudos Poliacias, Caxambu (MG).
- Hall, Stuart. 2001. "Constituting an Archive." *Third Text* 15 (4): 89–92. doi:10.1080/09528820108576903.
- Jovchelovitch, Sandra, and Jacqueline Priego-Hernández. 2013. *Underground Sociabilities: Identity, Culture, and Resistance in Rio de Janeiro's Favelas*. Brasília: UNESCO.
- Leite, Maria Periera. 2012. Da "Metáfora da Guerra" ao Projeto de 'Pacificação': Favelas e Políticas de Segurança Pública no Rio de Janeiro". *Revista Brasileira de Segurança Pública* 6 (3): 200–213.
- Lucas, Peter. 2013. *Viva Favela: Photojournalism, Visual Inclusion, and Human Rights in Rio de Janeiro*. Self-published, New York, New York.
- Mackay, Rhona. 2016. "Exposição 'Favela Tem Memória' Traz ao Público a Memória Social das Favelas do Rio". Rio Olympics Neighborhood Watch, March 16, 2018. <http://rioon-watch.org.br/?p=18728>
- Matheus, Leticia. 2011. *Narrativas do Medo: O jornalismo de sensações além do sensacionalismo*.
- Mayer, Vicki. 2006. "A Vida Como Ela é/pode Ser/Deve Ser? O Programa Aqui Agora e Cidadania no Brasil". *Intercom: Revista Brasileira de Ciências da Comunicação*. 29 (6): 15–37. <http://www.portcom.intercom.org.br/revistas/index.php/revistaintercom/article-view/38/1182>
- Meiskins Wood, Ellen. 1991. *The Pristine Culture of Capitalism: A Historical Essay on Old Regimes and Modern States*. New York: Verso Books.
- Mignolo, Walter. 2012. *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking*. 2nd ed. Princeton, NJ: Princeton University Press.
- Monteiro, Fabiola. 2016. Memória do Estácio. "Favela Tem Memória." <http://favelatemmemoria.com.br/samba-o-ontem-o-hoje-o-amanha-memoria-do-estacio/>
- Mpofu, Shepherd. 2014. "When the Subaltern Speaks: Citizen Journalism and Genocide 'Victims' Voices Online." *African Journalism Studies* 36 (4): 82–110. doi:10.1080/23743670.2015.1119491
- Perlman, Janice. 1980. *The Myth of Marginality: Urban Poverty and Politics in Rio de Janeiro*. Berkeley, CA: University of California Press.
- Perlmann, Janice. 2010. *Favela: Four Decades of Living on the Edge in Rio de Janeiro*. Oxford and New York: Oxford University Press.
- Pinho, Patricia. 2004. *Reinvenções da África na Bahia*. São Paulo: Editora Annablume.

- Platt, Damian and Patrick Neate. 2004. *Culture is our Weapon: Making Music and Changing Lives in Rio de Janeiro*. London: Penguin Books.
- Rabinowitz, Paula. 1994. *They Must Be Represented: The Politics of Documentary*. New York: Verso Books.
- Rodriguez, Dylan. 2007. "The Political Logic of the Non-Profit Industrial Complex." In *The Revolution Will Not Be Funded: Beyond the Non-Profit Industrial Complex*, edited by Incite, 2–19. Boston, MA: South End Press.
- Scott, Joan W. 1991. "The Evidence of Experience." *Critical Inquiry* 17 (4): 773–779.
- Soares, Luis E. 2006. "Segurança Pública: Presente e Futura." *Estudos Avançados* 20 (56): 91–106. doi:10.1590/S0103-40142006000100008.
- Tactical Technology Collective. 2009. *10 Tactics for Turning Information into Action*. London: Tactical Technology Collective.
- Thurman, Neil. 2008. "Forums for Citizen Journalists? Adoption of User-Generated Content Initiatives by Online News Media." *New Media & Society* 10 (1): 139–157. doi:10.1177/1461444807085325
- Thrall, Trevor, Dominik Stecula, and Diana Sweet. 2014. "May We Have Your Attention Please? Human-Rights NGOs and the Problem of Global Communication." *The International Journal of Press/Politics*. 19 (2): 135–159. doi:10.1177/1940161213519132
- Universo Online (UOL). 2018. "Adolescente é morto durante un tiroteio na Maré. Avenida Brasil e linhas Vermelha e Amarela são fechadas." *Vida Cotidiana*, February 6, 2018. <https://noticias.uol.com.br/cotidiano/ultimas-noticias/2018/02/06/rio-tem-linha-vermelha-linha-amarela-e-avenida-brasil-fechadas-apos-morte-de-adolescente-na-mare.html>
- Viva Rio. 2009. "Favela Tem Memória". [www.vivario.org.br/favela-tem-memoria](http://www.vivario.org.br/favela-tem-memoria)



# @FRANKLINFORDBOT

## Remediating Franklin Ford

**Juliette De Maeyer** and **Dominique Trudel**

*Franklin Ford (1849–1918) is mostly known for his association with the philosopher John Dewey in the late 1880s and early 1890s. Together, they attempted to launch Thought News, a “philosophical newspaper” that never saw the light of day. But both before and after that failed project, Ford never stopped developing a vision for the future of the news. Reading Ford is a jumping-off point for experimentations that raise original methodological questions in the field of media history and theoretical developments that speak to contemporary media problems. In that regard, our paper focuses on the methodological experiment undertaken to explore Ford’s work: the creation of an automated Twitter account, a “bot” that uses text-mining techniques to automatically tweet excerpts from his writings. The paper describes the concrete steps of that remediation: from the delineation of Ford’s written work to the gathering and digitization of the material and its transformation into tweetable soundbites. We argue that this combination of close and automated reading offers heuristic elements of surprise to guide the historical inquiry. As the tweets echo the specific genre of today’s “future-of-the-news” thinkers, they also constitute an attempt to explore the relationship between “old” and “new” media.*

Rather than merely creating a product called content and attracting an audience to sell to advertisers—our old model—we can now reconceive of journalism as a service to our communities, convening them into informed, civil, and productive conversation and helping them improve their lives – Jeff Jarvis (2017)

Like conversation the news business classifies according to relationship. Journalism, the registration of life through newspapers, leaflet and book, is but conversation writ large – Franklin Ford (1892)

Even as they evoke converging themes (the news business, conversation), these two quotes are separated by more than a century. The first is excerpted from a post published in August 2017 on *Medium* by Jeff Jarvis, a blogger and j-school Professor who stands prominently among what can be called the “future-of-news” thinkers (Starkman 2011). The second is from *Draft of Action*, a self-published document authored in 1892 by Franklin Ford, a Michigan-born journalist and editor who was, in his own way, also a “future-of-news” thinker, a media theorist, and an entrepreneur. What

brings these two characters together is a certain prophetic verve, a deep conviction that the news ecosystem needs a thorough reform of its economic model, and a fascination for the changes brought about by new technologies (be they networked platforms or the telegraph).

If future-of-news thinkers occupy a central place in contemporary scholarly debates, the very nature of their work, which emphasizes novelty and technological breakthroughs, obfuscates that these questions are as old as journalism, and that there are numerous histories of the future of the news, of “news-as-conversation,” and the news business that remains to be written. In this paper, we argue that digital technologies can contribute to the exploration and writing of these histories by opening alternative ways to analyze and disseminate archival material. To this end, we will detail a methodological experiment that is embodied in the creation of an automated Twitter account named @franklinfordbot (<https://twitter.com/franklinfordbot>), a “bot” which allows to “remediate” the work of Franklin Ford.

The first part of this paper offers an overview of the historiographical issues raised by the work of Franklin Ford. We then argue that what Ford calls the “movement of intelligence,” a theory about the circulation of facts in the media ecosystem, can be revisited in the light of digital technologies (namely Twitter bots). This combination of old and new media is approached through the lens of recent development in media history and in media archaeology, and it can be understood as an operation of (retro)-remediation. The second part of this paper describes the concrete steps of the series of remediations that is @franklinfordbot, and shows how they guided our historical inquiry.

### **Who is Franklin Ford and Why Does He Matter in the History of Journalism and Communication Research?**

Ford has left a mark in the history of communication research for what is mostly remembered as a brief supporting role in an obscure yet important episode. Between 1888 and 1892, at the University of Michigan at Ann Arbor, Franklin Ford planned to launch a revolutionary “philosophical” newspaper, called *Thought News – A Journal of Inquiry and a Record of Fact*. Ford’s main partners in the project were his brother Corydon as well as the philosopher John Dewey. Robert Ezra Park, Charles Horton Cooley, and George Herbert Mead were also involved in the project. All of them were in the early years of careers—Park and Cooley were still students—that would lead them to become prominent philosophers and social scientists. In 1892, the launch of *Thought News* was advertised in the local press but the “philosophical newspaper” never saw the light of day. Soon after Ford left Ann Arbor and the collaboration with Dewey, Park, Mead and Cooley seems to have stopped, with the possible exception of Park (Raushenbush 1979). Dewey’s correspondence portrays a very close relationship gone awry between the philosopher and the Ford brothers (Dewey to William James, June 3, 1891, cited in Perry 1935, 517–519; see also Martin 2002, 124–131).

*Thought News* has attracted the attention of numerous media and communication scholars (Carey and Sims 1976; Czitrom 1982; McGlashan 1976; Peters 1989; Carey 1989; Schiller 1996), who see it as an important episode in the early developments of the field. According to Carey (1989), “Research and scholarship on communication

began as a cumulative tradition in the United States in the late 1880s when five people came together in Ann Arbor, Michigan" (110). While Dewey, Park, Mead and Cooley were about to become key figures in the social sciences and the humanities in the United States, the same cannot be said of Franklin Ford, about whom very little is known besides his role—usually cast as minor—in the failed *Thought News* project. Ford, however, is the central character of this story. It is Ford who was the theoretical and conceptual kingpin of *Thought News*, as attested by a self-published, 58-page document titled *Draft of Action* (Ford 1892), in which he laid out the wider vision behind *Thought News*, a general theory he named "the movement of intelligence." In this respect, the "philosophical newspaper" was only a small part of Ford's grand plan, which also involved a complete reform of how the news was to be collected, processed and disseminated across the United States. While Ford's larger argument about the future of the news business remains little discussed by most scholars, the *Thought News* episode is believed to have had a lasting influence on Dewey, Park, Mead, and Cooley (Rauschenbush 1979; Martin 2002), and especially on Dewey's conception of communication, that remains central in the field of media and communication studies today (Westbrook 1991; Stroud 2011).

If Ford's ideas were crucial to the development of journalism and communication studies and still bear contemporary significance in the context of today's debates about the future-of-news, they still remain at the margins of media and communication historiography. Our project proposes a two-fold response to this deficiency, in the perspective of a contribution to the history of the future of the news, the history of the news business, and the history of "conversational" journalism. On the one hand, we aim at documenting Ford's little-known activities and writings that occurred before and after the *Thought News* project in order to recast Ford's work in its larger historical context and intellectual environment. On the other hand, we are seeking to reassess and recirculate Ford's writings to audiences that are potentially interested in the questions he raised—including concerns about the future of the news, the effects of new technologies on the circulation of information, and the democratic capacities of the news media. Central to our project is the creation of an automated Twitter account, a "bot," which brings together the two objectives of this project and simultaneously constitutes a research method and a research output. In order to contribute to our historical inquiry and to recirculate Ford's writings, the bot aims at putting Ford's theory of the "movement of intelligence" into action, that is to say to use the "new" technology that is Twitter to circulate facts to create meaning and knowledge. This approach, as we will explain, is informed by the theoretical developments surrounding the concept of "remediation," which aims to problematize the complex entanglements of old and new media technologies and the constant oscillation between immediacy and hypermediacy that is central to digital media (Bolter and Grusin 1999).

### **The "Movement of Intelligence"**

In its most concrete guise, Ford's program for the future of the news takes shape as a centralized news system that involves a triple distribution of facts (what Ford calls the "intelligence triangle"), thanks to three types of publications: general-interest newspapers, "class interest" newspapers that cater for the needs of specific professional communities

(Ford's planned to launch a series specialized publications bearing titles such as *Cotton, Grain, Fruits, Chemicals*, etc.), and an information bureau that sells customized facts to individuals. Ford went into many details in describing the nitty gritty of the system, complete with a list of geographic locations, names for the newspapers, and a business model (Ford 1892). Ford further developed the scheme during his entire life, both in his correspondence and in a series of little-known publications. For example, in a letter to his friend and Columbia University Librarian James H. Canfield sent on February 11, 1907, Ford enclosed further details, including a list of potential subscribers and a plan for incorporation. On another occasion, Ford published a five-page booklet, *Government is the Organization of Intelligence or News* (1905), in which he describes how "the news movement will become the primary influence in municipal affairs."

The overarching principle that sustains the model is what Ford calls the "movement of intelligence" or the "news movement," that is, the adequate flow of information through society (which Ford evokes in terms of the "social organism"). According to this principle, each "physical fact" is to be delivered to the appropriate audience by the appropriate medium and at the appropriate rhythm and time. The movement of intelligence may seem like a simple proposition to balance the informational supply and demand, by delivering just the right amount of facts to the right people. But it is more than just the fine-tuning of news distribution. Thanks to new technologies such as the telephone or the telegraph, Ford's ambition was also to make society aware of itself through "conversation." Under the new organization of intelligence, citizens would "report their facts," eventually making every town "self-reporting" (Ford 1905, 3). In other words, what Ford envisions is a self-regulating, self-governing news system, ultimately leading to the disappearance of any form of government, as the social organism would then perfectly regulate itself. In the future, it would be "news as government" (*Janesville Daily Gazette*, February 12, 1907). Ford's movement of intelligence therefore cannot be reduced to a reform of the newspaper business. It also implies an overall transformation of society under new technological conditions, with deep consequences for governance and democracy. Ford's interest was no less than "deciphering the effect of modern communication on the organization of the State as a whole" (Ford 1905, 5).

Our methodological experiment aims at taking advantage of a new medium (Twitter) to revisit the core principles of Ford's "movement of intelligence" and remediate Ford's oeuvre. In other words, the "movement of intelligence" is simultaneously the content and the overarching principle of the experiment in remediation that is the object of this paper.

### **Media History and Twitter Bots: An Experiment in Remediation**

Our methodological approach inherited from various recent developments in media theory and history. On the one hand, the "new history" of communication research (Pooley 2008; Thibault and Trudel 2015) aims to reevaluate deeply entrenched canonical narratives, notably those of the birth of communication research after the Second World War, and to approach the history of the field with great seriousness, interpreting archival material according to the highest standards of historical research. On the other hand, a different stream of research known as "media archaeology" (Parikka 2012; Zielinski 2006) problematizes the historicity and materiality of media

technologies. As historical knowledge depends on media for recording and transmission, media history is fundamentally conceived as a communication (and media) problem (Peters 2008). To explore this problem, media archaeologists have designed experiments that couple “new” and “old” media, reflecting on (and rehabilitating) their specific epistemologies and materialities. In doing so, media archaeology is not only concerned with writing new historical narratives, but also in “excavating the past in order to understand the past and future” (Parikka 2012, 2).

Drawing on both of these recent developments, our approach is also guided by the conceptual developments surrounding the concept of “remediation” (Bolter and Grusin 1999), that is, the operation through with the so-called “new” media refashion older media, or the act of incorporating previous media in a new medium. Using this concept helps us to design our experiment and has important epistemological and methodological implications. It implies that media history is a process through which “new” media transform older media forms, retaining some of their features and abandoning others. In other words, media history is an ongoing process of “refashioning” (Bolter and Grusin 1999).

Remediation also works the other way around, a process Bolter and Grusin (1999) called “retrograde remediation.” In order to “survive” a new mediascape, older media refashioned themselves and become more timely by virtue of discovering the “new” in the “old” (see Jutz 2011). In this sense, the automated Twitter bot aims at “(retro-)remediating” Ford’s “movement of intelligence” by refashioning this nineteenth-century media protocol (disguised as a Twitter bot) and by circulating Ford’s work to the appropriate audience, thanks to the appropriate medium, at an appropriate pace, through the appropriate technology.

We purport that Twitter is the ideal medium to “remediate” Ford’s ideas (and that Ford’s ideas is the ideal medium to “retroremediate” Twitter) for two distinct yet closely related reasons. First, Twitter has become a major platform in the rapid dissemination of the news, creating “personal publics,” that differ from the public of mass media in that “information is selected by criteria of personal relevance for a known, networked audience in a conversational mode” (Schmidt 2013, 7) In that respect, Twitter echoes Ford’s vision for the targeted dissemination of news in which notions such as verification, objectivity of fact-checking have very little importance. What matters is that the news reaches its intended audience at an optimal timing. Twitter is also a platform where citizens can “report their facts”: it has been characterized as an “ambient news network,” where users are part of the flow of news (Hermida 2010). In similar fashion, Ford writes, “the reporting machinery (...) is primarily the social organism itself. The citizen king is the crop reporter” (1892, 12). In these respects, the movement of intelligence and Twitter are both part of a culture of remediation that seeks to make temporal and technical mediation disappear in favor of apparent immediate and direct communication. If Ford’s “movement of intelligence” supposes an important technological machinery and media infrastructure, it is paradoxically to animate the organic movement of social life. In the case of Twitter, a similar trick consists in making individual voices central while obfuscating the multiple technical media layers supporting these voices: algorithms, protocols, network infrastructure, and so on. Therefore, our methodological experiment presents itself as a remediation of a remediation. Second, Twitter is home to many conversations that echo Ford’s preoccupations, such

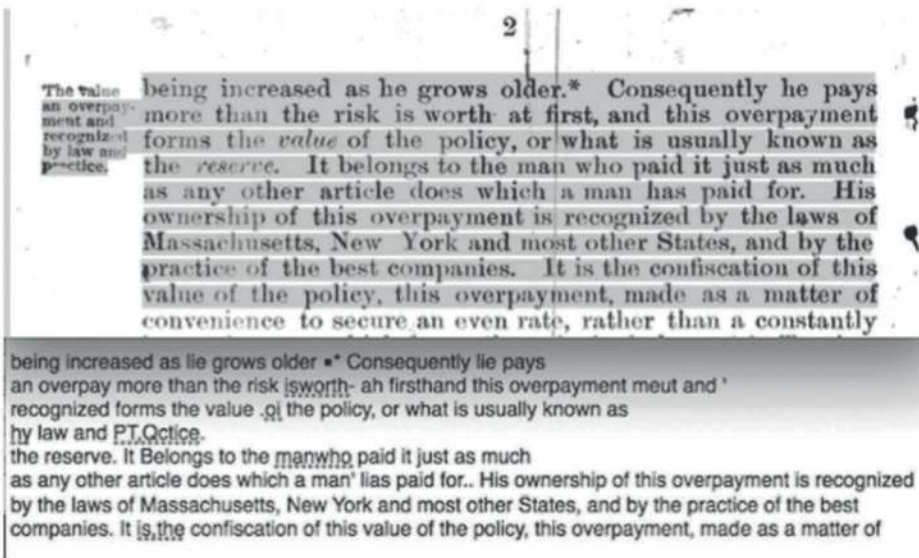
as the future of the news and the role of technology in our media ecosystem. It is also central to contemporary debates about “conversational journalism,” a concept nowadays associated with conversational bots that are not very different from @franklinfordbot. A conversational Twitter bot is the best means to enter the ongoing conversation about conversational Twitter bots.

In sum, following an archaeological impetus, our project seeks to (retro)remediate a nineteenth-century theory (and media protocol), the “movement of intelligence,” and a twenty-first-century technology, Twitter. The remainder of this paper specifically addresses the following questions: What are the concrete steps of this process of (retro)remediation? Can the process of building a Twitter bot better draw attention to the various material layers through which history is (always) mediated? How does it guide the historical inquiry in order to better understand the contours of Ford’s contribution to communication and journalism history?

In the next sections, we describe the concrete steps of this remediation process. It implies to juxtapose ways of writing that belong to radically different realms of existence: the heterogeneous material that constitutes Ford’s oeuvre (letters, pamphlets, news articles, leaflets, and books produced between 1874 and 1910), the API of Twitter as well as text-mining and automated publishing algorithms written in Python, a programming language. In doing so, we place side by side media layers that are incongruent with each other, an analytical move that underlines how, both in history and in media studies, interpretation always happens “under conditions of remoteness and estrangement” (Peters 2008, 40). Such juxtapositions are not a gratuitous way to multiply confrontations between old and new media, they also take advantage of the productive tension between “close” and “distant” reading (Bode 2017) and offer alternative entry points into the archival material. “Distant reading” (Moretti 2013) is the term used by digital humanities scholars to describe analyses of literary corpus that rely on aggregated data and computational methods, understanding texts as something that can be modeled, processed, and measured—as opposed to the “close reading” that characterizes humanities scholarship in its fine-grained relation with texts, documents and/or archives. In our case, it is not the scale of data that demands the assistance of machines—Ford’s written oeuvre is not exactly big data—but the randomized intervention of the bot (described below) that reorganizes the source material in original and productive ways. The tweets act as heuristic elements of surprise, highlighting parts of Ford’s thinking that had eluded our attention, driving some points home, or offering transversal pathways into the material. As we will show, the tweets accompanied us in our historical inquiry, therefore actively shaping our understanding of the historiographical problem discussed above, that of the contours of Ford’s contribution in its larger historical context and intellectual environment. In other words, the Twitter bot is not only an instrument of research output, a way of publishing results to an audience that differs from that of an academic journal, but it is also an integral part of the historical inquiry that emphasizes a nonlinear perspective.

### **Anatomy of a Bot**

To become tweets by @franklinfordbot, the archival records need to undergo several transformations. First, the various documents composing the archives have



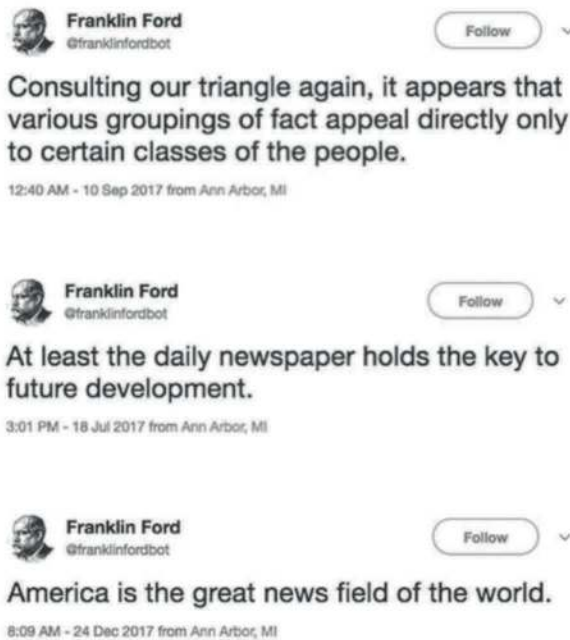
**FIGURE 1**

An excerpt from *Tontine* (Ford, 1882) transformed by OCR

been digitized, either by us or by the libraries that hold them (as part of larger digitization policies, or at our request). This first transformation, from paper to PDF, implies that “documents are experienced as pictures of themselves” (Gitelman 2014, 114). When possible, we transformed the page images into digital text, thanks to optical character recognition (OCR) software. Each document then becomes a string of characters that can be computationally processed. Those strings of text are augmented with the relevant metadata (source, place of publication, and date) indebted to our close reading of the archival material.

There are two important transformations at work here: first, OCR softwares are far from perfect, especially in the case of historical documents (Milligan 2013), producing digital texts that cannot be understood as the exact reproduction of the original documents: OCR engines “compose” as much as they read documents (Cordell 2017). Figure 1 shows how an excerpt of Ford’s *Tontine: What It Is; How It Works* (1882) has been transformed by OCR, which in this case produces a considerable amount of noise. Second, for the purpose of processing, the documents are stripped of anything that is not textual (layout, images, placement, texture, preservation or degradation, etc.), therefore losing all the important material clues related to the “conditions of encoding” (Peters 2008, 21) that are at the core of a historian’s assessment of a source.

The text is then broken down into sentences, thanks to the sentence segmentation function of NLTK, a natural language processing toolkit (Bird, Klein, and Loper 2009). This operation is comparable to previous experiments in media history, notably what Ferguson (2016) calls “slicing,” that is the automated, indiscriminate transformation of “media text into something wholly new as an object of investigation by first cutting it into pieces” (275). The result of the slicing is a list of sentences, which is then filtered according to a list of that we have manually determined during our close reading of the texts. Our list of 67 keywords has been prepared based on what we then thought were the fundamental historiographical issues at stake and the most



**FIGURE 2**  
Tweets with the largest engagements

interesting topics addressed by Ford. For example, our keywords aim at characterizing the intellectual influences of Ford ("Tarde," "Darwin," "Trotter," etc.), at mapping his social network and professional trajectory ("Vail," "Bradstreet's," "Columbia," etc.), and at exploring the metaphor of the "social organism" ("body," "ganglions," etc.). Finally, other keywords broadly referred to our research object and theme of the bot ("conversation," "journalism," "news," etc.). Another approach would be to calculate TF-IDF ("Term Frequency-Inverse Data Frequency") scores for all the words in the corpus in order to generate a list of keywords based on occurrence (see Wang et al. 2015). We consider adopting this strategy at a later step of the project.

This list of filtered sentences (and their attached metadata about the source, date, and location) is the raw material used by the bot itself: a Python script that randomly chooses a sentence among the list of all possible sentences, measures the length of the sentence, divides it into "tweetable" pieces (i.e., 140 characters or less) and publishes them on Twitter at a random interval that ranges between 18,000 and 176,400 s (i.e., 5 h and 49 h, those limits have been chosen arbitrarily). If a sentence's metadata contains a location, such as the publication place, the script transforms it into latitude/longitude coordinates (thanks to a manually-determined list of correspondences between locations and coordinates) and adds this as the geolocation of the tweet.

@franklinfordbot started to tweet on March 21, 2017. Between then and the time of writing (May 27, 2018) it has published 537 tweets and gained 54 followers. According to the analytics provided by Twitter, those 537 tweets gained a total of 91,954 "impressions," that is, the number of times users saw the tweet, and 894



“engagements” (total number of times a user has interacted with a tweet, including all clicks anywhere on the tweet, retweets, replies, follows and likes), 60 retweets and 51 likes. Figure 2 shows the three tweets that have gained the highest engagement.

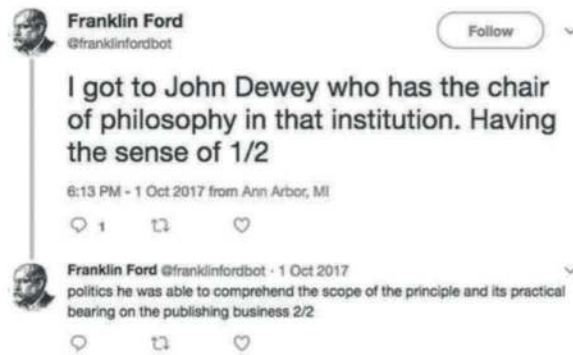
### Locating and Circulating Ford’s Work

Locating Ford’s written works turned out to be a challenge. In 1914, a fire that wrecked the commons and gymnasium of Columbia University at least partly destroyed them. A report printed in the *New York Tribune* lamented that “papers representing the work of twenty years are believed to have been ruined” (*New York Tribune* 1914, 10) in his office—Ford occupied various rooms at Columbia University Library, starting in 1907, thanks to his acquaintance with head librarian James Canfield. His contributions to newspapers, during his time at the *Baltimore Gazette*, the *Philadelphia Record*, the *New York Sun* or during his time as an editor of *Bradstreet’s—A Journal of Trade, Finance, and Public Economy* are not bylined, in accordance with the standards of the late nineteenth century press. What is left of Ford’s printed work therefore mostly exists in the obscure realm of self-published books and opuscles, reports of talks given at conferences or club meetings, and a rich correspondence with various people Ford wanted to enroll in his projects.

Our efforts to gather Ford’s written work therefore resembled an archival puzzle that started from the fragments that have been under scrutiny, notably among those who have worked on the *Thought News* episode, such as *Draft of Action* (Ford 1892). Large-scale digitization efforts undertaken by various actors such as Google Books, public libraries, the Internet Archive, Newspapers.com, or Hathi Trust may give the impression that archives are within reach, only a few clicks away. But our first attempts at discovering other pieces of Ford’s works (by naively typing “Franklin Ford” in various search engines) remained mostly fruitless, an example of the “illusionary order” of online databases (Milligan 2013). The absence of results did not mean, however, that there was nothing by Ford or about Ford in those databases, only that the ubiquitous “search” function cannot be taken for granted (Robson 2015): algorithmic explorations in the form of full-text search comes with technical limitations—such as the “errorful” mediation of optical character recognition software (Cordell 2017)—as well as interpretative limitations, that is, a specific ordering of documents and the hermeneutic assumption that one knows what they are looking for (Underwood 2014).

### From the Archives to Tweets and Back Again: Strolls between Old and New Media

Instead of an ordered process of search and results, we therefore proceeded in a series of back-and-forth movements between the “distant reading” of databases, algorithms, and tweets from @franklinfordbot, and the “close reading” of archival documents that are at the core of historical research—acting as what Robson (2015) calls “double agents,” conducting our inquiry in “both newfangled and oldfangled ways” (13). The following section describes the kind of documentary pathways that we explored thanks to this approach. It is only an example of one itinerary among others, but it illustrates how iterations, detours and leaps in our heterogenous material as well



**FIGURE 3**

Ford got to John Dewey

as the various operations of remediation of @franklinfordbot allowed us to highlight aspects of Ford's life and contributions that were previously unexplored.

Two seemingly unrelated tweets by @franklinfordbot highlighted specific questions related to Ford's career in the 1880s, before the *Thought News* episode, when he was the editor of *Bradstreet's*. The first tweet (Figure 3) evokes the moment when Ford "got to John Dewey." The second tweet (Figure 4) seems to indicate that Ford, as the editor of *Bradstreet's*, received queries from businessmen to provide personalized news reports. These tweets, obviously, do not have the effect of leading directly to new extraordinary discoveries. The question of Ford's professional trajectory, in the context of this research, arises from the outset, and the question of the nature of his relationship with Dewey is already central in the existing historiography. But taken together, the tweets brought about a series of questions about the social status of Ford in the 1880s, in relation with his tenure as the editor of *Bradstreet's*, and they point to news-as-business or "news-as-goods" as an important object in this story—something that clashes with the "philosophizing" approach to *Thought News* that is generally adopted. Let us be clear: the juxtaposed tweets helped us to identify research leads that could possibly have been discovered otherwise, through a traditional close reading of the archival material. In this respect, our method does not supersede traditional historical methods. It is only a slightly different way to approach the archival material, with its own benefits and its own problems. Instead of a linear (and often chronological) reading of the archival material, our approach insists on shortcuts, juxtaposition, and the non-chronological exploration of different material strata (archives, digitized archives, OCR-read digitized archives, tweets, comments, etc.).

A search query for "Franklin Ford" in the Hathi Trust digital library listed, among other results, a book titled *Notable New Yorkers of 1896–1899* (King 1899). The book is a collection of portraits of important New Yorkers, and includes a photograph of Ford—until then, the only known picture of Ford was an engraving printed in the *Detroit Evening News*, alongside an article about the wedding of Ford and Mathilde Coffin (Carey and Sims 1976). But *Notable New Yorkers* also leads us to another book by the same author, King's *Handbook of New York City* (1893), a 928-page illustrated catalog of "every notable institution" in New York City. It contains a couple of pages on *Bradstreet's*, the company for which Ford worked as an editor between 1880 and 1887,



**FIGURE 4**

Ford solicited by businessmen seeking information

complete with an engraving representing the facade of its building, located on 279, 281, and 283 Broadway. Retracing this extremely modest documentary path—from a database to a book to another book—not only gave us iconographic material, but also rich information about the position and importance of *Bradstreet's* at the time.

We have thus refocused our research on Ford's career at *Bradstreet's*, during which Ford truly made a name for himself. In addition to his editorial responsibilities, Ford became acquainted with numerous journalists, politicians and intellectuals, and was a member of the City Club and the nineteenth-century Club. In 1882, he published *Tontine: What It Is; How It Works*, a pamphlet arguing that the Tontine insurance scheme, then on the rise, was a fraud. The following year, Ford was involved in the debates surrounding New York Mayor Edson's project to reform the city charter, and published a short collection of his observations. Ford's expertise in municipal affairs eventually led him, in 1886, to be invited by New York Mayor Grace to join a committee in charge of crafting recommendations to the New York State Constitutional Convention to achieve greater municipal autonomy. Alongside with Ford, future Secretary of State and Nobel Prize winner Elihu Root and influential New York Lawyer Wheeler H. Peckham participated in the 12-man committee.

Not only do those iterations between tweets and databases point to new archival records (such as the papers of Elihu Root and Wheeler H. Peckham, both at the Library of Congress), they also open up new vistas on Ford's social network as well as on his credibility: some scholars have tended to characterize Ford as a "scoundrel"—the word comes from Dewey himself (Martin 2002, 135)—"a sort of crackpot journalist-philosopher" (Peters 1989, 253) who managed to enroll John Dewey and others into his unrealistic plans. But these documents tend to show that Ford was the respected editor of a notable institution, which makes the fact that he "got to John Dewey" less of an accident, and more of the result of Ford's background and social trajectory.

Focusing on *Bradstreet's* also illuminates the genealogy of Ford's will to reform the news. In *Draft of Action* (1892), Ford notes how he started to think about the "movement of intelligence" while he was the editor of *Bradstreet's*, in 1883. More about these early developments, almost a decade before *Thought News*, is to be found in the correspondence between Ford and Edward Atkinson, an industrialist and activist for various social causes, whose papers are held at the Massachusetts Historical Society.

Interestingly enough—and another example of the fuzziness of searches in online databases—the finding aid of the Atkinson papers identifies the editor of *Bradstreet's* under the name “Franklin L. Ford,” whereas Ford’s full Christian name is “William F. Ford” and most of his work known under the name of “Franklin Ford.”

Ford’s letters to Atkinson reveal that as early as 1886, he suggested to the President of *Bradstreet's*, Charles F. Clark, that he completely reorganizes the operations of *Bradstreet's*. Ford proposed to set up a publicity bureau supplying country papers with business and city news, to launch specialized papers (called Food, Metal, and Textiles), and to provide reports according to the specific interest of customers (Ford to Atkinson, October 3, 1886). This shines a new light on why and how Ford, as the editor of *Bradstreet's*, came to be solicited by businessmen to fulfill their information needs (an intriguing element that had first appeared in @franklinfordbot’s tweets—see Figure 4). Those customized requests were directed to an important institution: as Carey and Sims (1976) noted, credit reporting agencies like Dun’s (f. 1841) and *Bradstreet's* (f. 1859) were the first large-scale national centralized information services in the United States. Ford’s vision was modeled on the functioning of these mercantile agencies, which employed stringers in every corner of the country, gathering information that needed to be processed by (and broadcast from) a central organ, and also distributed to customers on a personalized basis.

The period Ford spent at *Bradstreet's* and how it shaped his will to reform the news also highlight how Ford’s contact with John Dewey and the others at the University of Michigan did not happen out of the blue. In 1887, after his ideas failed to gain traction at *Bradstreet's*, Ford toured the country in search of allies. In a letter to Edward Atkinson, Ford describes his undertaking: “I have got thus far in the work of visiting the chief intelligence centers. I have come from Chicago by way of St Paul, Omaha, Cheyenne, Denver, Leadville, Kansas City, St. Louis, Memphis and Nashville. I go from here to Galveston. From there I shall return to New York by way of Birmingham, Atlanta, Savannah and Charleston. I have succeeded in banding together the leading newspapers to receive intelligence from New York.” In May 1887, Ford was back in New York. According to a letter he sent to Edward Atkinson, October 11, 1887, Ford then tried to launch the operations of “Ford’s News” with the help of three associates, Walter H. Page, Frank W. Rollins, and Lindley Vinton. However, it seems that the editors Ford met had changed their minds. According to Ford, “the crush of fact” was not welcome among editors, and in order to find people genuinely interested in “the principle of intelligence,” he turned to the universities (1892, 2–3). The *Thought News* episode is therefore not an isolated incident, but rather a continued effort from Ford, started in the 1880s and pursued until Ford’s death in 1918.

### **Discussion: Twitter Bots as Tools for Remediation**

Our aim with creating @franklinfordbot was to use it both as a research method and a research output. On the one hand, it is an experiment with historical methodologies that emphasize the juxtaposition of media layers that are incongruent with each other. On the other hand, it is an embodiment of Franklin Ford’s “movement of intelligence,” that is, the adequate circulation of information to the appropriate audiences, as remediated by

Twitter. In sum, our Twitter bot “remediates” Franklin Ford in two different ways: not only it recirculates his work, it also extends his methodological precepts.

As a methodological experiment, @franklinfordbot successfully fed our documentary exploration of the archive. The slices of Ford’s oeuvre, selected and published randomly, were part of iterations between heterogeneous documents which, taken together, allowed us to open original vistas on Ford’s life and contribution to the development of media and communication research. As such, the bot contributes to experimentations in media history that emphasize non-linearity and the materiality of media, in the tradition of media archaeology. Moreover, it also converges with attempts to conceptualize “digital surrealism” as a method for media history (Ferguson 2016). Those approaches aim at revealing aspects of media texts that would not be visible without computer-assisted techniques, such as automated “slicing.” Just like the cut-up method or Salvador Dalí’s “paranoiac-critical method,” the randomized interventions of @franklinfordbot were instrumental in producing “irrational knowledge that springs from unexpected juxtapositions of unrelated elements” (Ferguson 2016, 274). The bot also echoes another key precept of avant-garde art, which has approached writing as an eminently collective—and sometimes automatic—act. One of the aims of the bot was not only to publicize our research but also to involve different communities, especially fellow scholars and students. Their comments, retweets and suggestions (along with the automatic work of the bot) are conceived as part of a collective methodological experiment in the field of media history. In this respect, our methodological experiment is conceived as a collective process of inquiry (or what Ford called a “movement”) that echoes the current attempts to use “crowdsourcing” in relation to archival material or in the context of emerging journalism practices such as “computational journalism” (Cohen, Hamilton, and Turner 2011). In this regard, an interesting aspect of Ford’s theory is to imagine journalism without journalists: infrastructure, networks, and the “citizen king” are the key elements to organize the collective social inquiry. Our methodological experiment is guided by similar ambitions, and although the movement is modest in scale, it has produced some results. For example, one descendant of a Ford’s associate (we voluntarily preserve her anonymity) contacted us to express her interest and gave us some interesting information.

As a practical remediation of the “movement of intelligence,” the bot managed to publish about 537 tweets to a relatively modest but specialized audience (54 followers, 91,954 “impressions” and 894 “engagements” is a rather small reach in the era of mass virality). In doing so, the algorithmic nature of the bot also meant to go beyond the core principle of the “movement of intelligence,” that is the fine-tuned, controlled and appropriate delivery of news to an audience, in a quest to balance informational supply and demand. Not every tweet made sense to its perfect audience, on the contrary, many of them showed how @franklinfordbot is a messy process with its (productive) failures, accidents and mishaps. For example, the different transformations necessary to turn Ford’s words into tweets, including the composition of OCR and the slicing into sentences, sometimes created tweets that were unintelligible or plainly uninteresting. Such unordered messiness is part of automation, and participates in the “digital surrealism” discussed above. It also shapes how the tweets are distributed and received by an audience. For example, the high number of total “impressions” (91,954) of the 537 tweets published between March 21, 2017, and May 27, 2018, comes partly



**FIGURE 5**  
Tweets with the most impressions

from three tweets published on the same day (April 1, 2017) that have each gained around 16,100 impressions—whereas the average number of impressions for the rest of the tweets is around 100. We do not have an exact explanation for these unusually high level of impressions, other than a possible action from Twitter’s algorithm that may have exposed those tweets—which are neither particularly interesting nor eloquent (see Figure 5)—to many users (or at least, to what algorithmically counts as “users”), unbeknownst to us and for reasons that will remain obscure. If anything, this glitch reveals the algorithmic, uncontrollable and hybrid nature of @franklinfordbot’s movement into the digital realm: the way the tweets circulate are partly due to how the followers interact with the content (notably by retweeting, which exposes the tweets to new audiences), but also due to algorithmic accidents.

Designed as an operation of remediation, our research was conceived as a series of “refashioning” operations: archival material was gathered and digitized, then read using an optical character recognition program, cut into sentences, etc. As the output of a whole series of remediations, the tweets are quite distinct from the original material. And although they present themselves as the voice of Franklin Ford, the tweets are the products of several operations and are in fact more akin to original creations.

This paradox—appearing as natural artifacts, but resulting from many operations of remediation—is central to contemporary media culture, which “wants both to multiply its media and to erase all traces of mediation: ideally, it wants to erase its media in the very act of multiplying them” (Bolter and Grusin 1999, 5). In the context of our research, we kept this tension between the unveiling and the erasure of the remediation operations: the sudden topical comments of @franklinfordbot, in the context of contemporary conversations about the future of the news, are both masking and showing a series of sedimented creative operations of remediation. The tweets are inside-out commentaries on the many remediations of historical inquiry (and a way to make these operations visible by dramatizing the historical inquiry), and at the same time, they point to contemporary events and debates. Among many examples, on December

24, 2017, @franklinfordbot tweeted “America is the great news field of the world” (see Figure 1). This tweet echoes Donald Trump’s famous motto to “make America great again” and his recurrent critical comments about the decay of mainstream media, somehow contradicting the two arguments.

This project’s contribution also speaks to the emerging issue of bots in democratic life (Woolley and Howard 2016). The term “bot” could be used to describe any kind of automated program, but recently has emerged to describe, more specifically, automated accounts on social media platforms such as Twitter or Facebook, also called “social bots” (Gehl and Bakardjieva 2016). Bots vary in scope and aim—weather bots, conversation bot, spam bots, poetic bots—but they are often discussed in terms of nuisances, political propaganda, and potential threat to democratic deliberation (McKelvey and Dubois 2017). In that perspective, bots can be used at a large scale to amplify or obscure information that circulates on social networking sites, or play a part in schemes that can involve identity theft, artificially swollen social media audiences, and lucrative frauds in the “influence economy” (Confessore et al. 2018). These works usually see automation as a threat and seek to clarify the blurring distinction between human and machine-produced contents and interactions online. Other lines of work have emphasized the creative capacities of bots, seeing automation as a form of artistic performance (Bucher 2014), a new avenue for the dissemination of the news with an interactive quality that could enhance user engagement (Barot 2015) or even a potentially democratic tool to support public deliberation (Graham and Ackland 2016) and the deployment of collective counterpublics (Geiger 2016). Our bot draws on those approaches that highlight the positive potential of automated interventions in public life: it is modeled on the “movement of intelligence” envisioned by Ford as a way of creating and funneling the optimal circulation of data and information in a democratic society (Pinter 2003).

Altogether, @franklinfordbot is still an open experiment that remains in progress. The tweets that have been published so far only cover a small part of Ford’s oeuvre, and there is enough material for the bot to keep on tweeting for several years. Furthermore, there are many ways in which we could enhance the bot to continue to experiment with Ford’s ideas: the conversational aspects of the “movement of intelligence” could lead, for example, to a bot that seeks to intervene in current conversations about specific topics (the future of the news, the role of media and technologies in democracy) that could be identified on Twitter via hashtags. Aspects of Ford’s theory that emphasize customization and personalized news could lead to the development of a bot that is more interactive (whereas it is currently limited to broadcast functions) and that could reply to queries that are addressed to it, for example by applying techniques of machine learning and artificial intelligence to not only tweet existing excerpts of Ford’s written work, but also to actively create new fragments—therefore also contributing to current research about conversational bots.

## ACKNOWLEDGMENTS

We are also grateful for the help we received from archivists at Columbia University Library, the Library of Michigan, the Massachusetts Historical Library, the AT&T Archives and History Center, the University of Washington Special Libraries, the Bentley Historical Library at the University of Michigan, and the

Delmar T. Oviatt Library of the California State University at Northridge. Finally, we also wish to thank the anonymous reviewers from Digital Journalism and the American Journalism Historians Association for their insightful and generous comments.

## FUNDING

This work was supported by the Social Sciences and Humanities Research Council [grant number 430-2018-00809].

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

## REFERENCES

- Barot, Trushar. 2015. "The Botification of News." *Nieman Lab*. <http://www.niemanlab.org/2015/12/the-botification-of-news/>.
- Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.
- Bode, Katherine. 2017. "The Equivalence of 'Close' And 'Distant' Reading; Or, toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78 (1): 77–106. doi: 10.1215/00267929-3699787.
- Bolter, Jay D., and Richard Grusin. 1999. *Remediation: Understanding New Media*. Cambridge: MIT Press.
- Bucher, Taina. 2014. "About a Bot: Hoax, Fake, Performance Art." *M/C Journal* 17 (3). <http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/814>.
- Carey, James W. 1989. *Communication as Culture*. Boston: Unwin Hyman.
- Carey, James W., and Norman Sims. 1976. "The Telegraph and the News Report." In Paper presented at the annual meeting of the Association for Education and Journalism, College Park, August.
- Cohen, Sarah, James T. Hamilton, and Fred Turner. 2011. "Computational Journalism." *Communications of the ACM* 54 (10): 66–71. doi:10.1145/2001269.2001288.
- Confessore, Nicholas, Gabriel J.X. Dance, Richard Harris, and Mark Hansen. 2018. "The Follower Factory." *New York Times*, January 27.
- Cordell, Ryan. 2017. "Q i-Jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20 (1): 188–225. doi:10.1353/bh.2017.0006.
- Czitrom, Daniel J. 1982. *Media and the American Mind: From Morse to McLuhan*. Chapel Hill: University of North Carolina Press.
- Ferguson, Kevin L. 2016. "The Slice of Cinema: Digital Surrealism as Research Strategy." In *The Arclight Guidebook to Media History and the Digital Humanities*, edited by Charles R. Acland and Eric Hoyt, 270–299. <http://projectarclight.org/book/>.
- Ford, Franklin. 1882. *Tontine: What It Is; How It Works*. New York: Self-published.
- Ford, Franklin. 1892. *Draft of Action*. Ann Arbor: Self-published.
- Ford, Franklin. 1905. *Government is the Organization of Intelligence or News*. New York: General News Office.



- Gehl, Robert W., and Maria Bakardjieva. 2016. *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. New York: Taylor & Francis.
- Geiger, R. Stuart. 2016. "Bot-Based Collective Blocklists In Twitter: The Counterpublic Moderation Of A Privately-Owned Networked Public Space." *AoIR Selected Papers of Internet Research* 5. <https://spir.aoir.org/index.php/spir/article/view/1076>.
- Gitelman, Lisa. 2014. *Paper Knowledge: Toward a Media History of Documents*. Durham: Duke University Press.
- Graham, Timothy, and Robert Ackland. 2016. "Do Socialbots Dream of Popping the Filter Bubble? The Role of Socialbots in Promoting Deliberative Democracy in Social Media." In *Socialbots and Their Friends: Digital Media and the Automation of Sociality*, edited by Robert W. Gehl and Maria Bakardjieva, 187–206. New York: Routledge.
- Hermida, Alfred. 2010. "From TV to Twitter: How Ambient News Became Ambient Journalism." *M/C Journal* 13 (2). <http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/220>.
- Jarvis, Jeff. 2017. "If I Ran a Newspaper ..." *Medium*. <https://medium.com/whither-news/if-i-ran-a-newspaper-220a065d2232>.
- Jutz, Gabriele. 2011. "Retrograde Technicity and the cinematic Avant-Garde: Towards a New *Dispositif* of Production." *Recherches Sémiotiques* 31 (1–3): 75–94. doi:10.7202/1027442ar.
- King, Moses. 1893. *King's Handbook of New York City*. Boston: M. King.
- King, Moses. 1899. *Notable New Yorkers*. New York: M. King.
- Martin, Jay. 2002. *The Education of John Dewey: A Biography*. New York: Columbia University Press.
- McGlashan, Zena Beth. 1976. "John Dewey and News." *Journal of Communication Inquiry* 2 (1): 3–14. doi:10.1177/019685997600200102.
- McKelvey, Fenwick, and Elizabeth Dubois. 2017. "Computational Propaganda in Canada: The Use of Political Bots." Computational Propaganda Research Project – Working Paper No. 2017.6. <http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Canada.pdf>.
- Milligan, Ian. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." *Canadian Historical Review* 94 (4): 540–569.
- Moretti, Franco. 2013. *Distant Reading*. London: Verso Books.
- Parikka, Jussi. 2012. *What Is Media Archaeology?* Cambridge, UK: Polity.
- Perry, Ralph. B. 1935. *The Thought and Character of William James, Vol. II*. London: Humphrey Milford/Oxford University Press.
- Peters, John Durham. 1989. "Satan and Savior: Mass Communication in Progressive Thought." *Critical Studies in Mass Communication* 6 (3): 247–263. doi:10.1080/15295038909366751.
- Peters, John Durham. 2008. "History as a Communication Problem." In *Explorations in Communication and History*, edited by Barbie Zelizer, 19–34. New York: Routledge.
- Pinter, Andrej. 2003. "Thought News a Quest for Democratic Communication Technology." *Javnost - The Public* 10 (2): 93–104. doi:10.1080/13183222.2003.11008830.
- Pooley, Jefferson. 2008. "The New History of Mass Communication Research." In *The History of Media and Communication Research*, edited by David Park and Jefferson Pooley, 43–69. New York: Peter Lang.
- Rauschenbush, Winifred. 1979. *Robert E. Park: Biography of a Sociologist*. Durham, NC: Duke University Press.

- Robson, Catherine. 2015. "How We Search Now: New and Old Ways Of Digging Up Wolfe's 'Sir John Moore.'" In *Virtual Victorians: Networks, Connections, Technologies*, edited by Veronica Alfano and Andrew Stauffer, 11–28. New York: Palgrave Macmillan.
- Schiller, Dan. 1996. *Theorizing Communication: A History*. New York: Oxford University Press.
- Schmidt, Jan-Hinrik. 2013. "Twitter And The Rise Of Personal Publics." In *Twitter and Society*, edited by Katrin Weller, Axel Bruns, Jean Burgess, and Cornelius Puschmann, 3–14. New York: Peter Lang.
- Starkman, Dean. 2011. "Confidence Game: The Limited Vision of the News Gurus." *Columbia Journalism Review*. [http://www.cjr.org/essay/confidence\\_game.php?page=all](http://www.cjr.org/essay/confidence_game.php?page=all).
- Stroud, Scott R. 2011. *John Dewey and the Artful Life*. University Park: Penn State University.
- Thibault, Ghislain, and Dominique Trudel. 2015. "Excaver, tracer, réécrire: sur les nouveaux historiques en communication." *Communiquer. Revue de communication sociale et publique* 15: 5–23.
- Underwood, Ted. 2014. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127 (1): 64–72. doi:10.1525/rep.2014.127.1.64.
- Wang, Xin, Neil T. Bendle, Feng Mai, and June Cotte. 2015. "The Journal of Consumer Research at 40: A Historical Analysis." *Journal of Consumer Research* 92 (1): 5–18.
- Westbrook, Robert. 1991. *John Dewey and American Democracy*. Ithaca: Cornell University Press.
- Woolley, Samuel C., and Philip N. Howard. 2016. "Political Communication, Computational Propaganda, and Autonomous Agent." *International Journal of Communication* 10 (2016): 4882–4890. <http://ijoc.org/index.php/ijoc/article/view/6298/1809>
- Zielinski, Siegfried. 2006. *Deep Time of the Media: Toward an Archaeology of Hearing and Seeing by Technical Means*. Cambridge, MA: MIT Press.



**Taylor & Francis**

Taylor & Francis Group

<http://taylorandfrancis.com>

# Index

- accuracy score 46–47  
actors 83, 117–118, 120–121, 166  
affordances 114–116; digital archival 115, 119;  
theory 114–115  
Agarwal, Sheetal D. 135  
agency 56, 110, 122–123  
age of profusion 18  
America 6, 53–57, 59–68, 70, 172  
Amin, Shahid 147  
Ananny, Mike 95  
Anderson 95–96  
archival material 112, 163–164, 167, 170–171  
archival processes 117  
archival research 2, 41, 48  
archived webpages 74–75, 85  
archives/archiving 2–8, 10, 14–16, 18–19, 39,  
53–54, 58–59, 68–70, 74–83, 87–88, 94–99,  
102–107, 110–123, 127–128, 131–138,  
143–155, 163, 166–167; and aspiration  
147; digital 7, 105, 119, 127–128, 134, 167;  
emulation-based 98, 104; established 5;  
experimental 5, 120; news apps 103; online  
news 94; oral history 113, 121; politics of  
5–6, 111, 115; process 10, 97–99, 103–104,  
106, 127; resource-depleted newsrooms 103;  
system 101, 107  
Atkinson, Edward 168–169  
Atteveldt, Wouter van 65  
  
Barnhurst, Kevin G. 11–12  
Barthel, Michael L. 135  
Bathran, Ravichandran 136  
Bickham, Troy 56  
Birkner, Thomas 12  
Birmingham Daily Post 65–67  
Blei, David M. 60  
Block, Sharon 58, 66  
Bob, Clifford 155  
Bolter, Jay D. 162  
Boumans, Jelle W. 43, 58  
Bourmans, Jelle W. 136  
Bowker, Geoffrey C. 95  
Boyd, Danan 20, 48, 87  
*Bradstreet* 165–169  
  
Braun, Joshua A. 95  
British India to India and Pakistan 113  
broadcast/broadcasting 5, 12, 24–34, 95, 169  
Broersma, Marcel 6, 38, 40  
Bryant, William Cullen 32  
Burscher, Björn 40, 44–45  
  
Carey, James W. 159, 169  
Carpentier, Nico 115  
Castells, Manuel 128, 131–132, 137–138  
cause-and-effect thinking 25  
Chopra, Rohit 130  
citizen journalism 127, 134–136  
communication research 96, 159, 161, 170  
communities 79–83, 85, 87–88, 100, 122,  
130, 134–135, 143–147, 149–153, 158, 170;  
marginalized 111, 143; members 143–144,  
147, 151  
completeness 4, 76, 79–81, 88  
content analysis 18, 40, 44, 47–48, 58, 82–83,  
95; automatic 15, 40, 43–45, 47  
Cooley, Charles Horton 159  
copyright 19, 102, 105–106  
corpus 4–6, 10, 45, 58–59, 63, 66, 68, 70, 83,  
96, 146  
Couldry, Nick 117  
counter-archives 110, 120, 123  
Crawford, Kate 20, 48, 87  
culture 10, 25–26, 54, 74–75, 134, 137, 146, 162  
customization 116, 119, 172  
  
Dahlgren, Peter 41  
Dalit archive 133, 135  
Dalit Camera (DC) 127–128, 130–138; digital  
archive of 137–138  
Dalit history and identity 130  
Dalit newspapers 129  
Dalits 127–138  
databases 39, 43, 77, 80, 85, 96–97, 105,  
112–114, 117–118, 121–122, 166, 168  
data collection 80, 82, 86–88, 122  
data journalism 6, 94–97, 103, 107; stories  
96–97, 99  
Deacon, David 39

- depth 47, 79–81, 88  
 digital archive affordances 115–117, 119  
 digital humanities 2–3, 60, 75, 120, 123, 144, 146  
 digitalization 9–10, 13–14, 16; of newspapers 11, 13, 15  
 digital journalism 2, 6, 8, 14, 54–55, 62, 70–71, 74, 95–96, 117, 120, 123, 136; studies 8, 70, 116, 127, 129, 136  
 digital media 111, 116, 134, 136–137, 144, 149, 153, 155, 160  
 digital research 15  
 digital technologies 2, 5, 7, 112, 128, 135, 147–149, 159  
 digitization 2, 4, 13, 19, 38–39, 75, 111–112, 158; of newspaper archives 38  
 Djerf-Pierre, Monika 114–115  
 Dominik Stecula, and Diana Sweet 155  
*Draft of Action* 168  
 Dynamic Data and Databases 77
- eclecticism 13, 16, 19  
 Eichhorn, Kate 111  
 emulation-based archiving tools 94, 104  
 emulation-based web archiving tool 103, 107  
 English and Foreign Language University in Hyderabad 127  
 Evans, Sandra 115  
*Event, Metaphor, Memory: Chauri Chaura 1922–1992* 147
- favelas 7, 143–145, 147–154  
 Favela Tem Memoria archive 145, 150–152, 155  
 Ferguson, Kevin L. 164  
 Fischer, Brodewyn 151  
 Ford, Franklin 7, 158–161, 166–169, 171  
 future-of-news thinkers 158–159
- Günther, Elisabeth 59  
 genre categories 46–47  
 genres 6, 41–47, 63, 158  
 Germany 5, 9–16  
 Ghersetti, Marina 114–115  
 Gibson, James 114  
*Global Elements of Literary Theory* 146  
*Government is the Organization of Intelligence or News* 161  
 Grusin, Richard 162  
 Guha, Ranajit 137
- Handbook of New York City* 167  
 Hanitzsch, Thomas 135  
 Hartsock, John 42  
 Hedman, Ulrika 114–115  
 Hermida, Alfred 48  
 historical analysis 24, 26–27  
 historical archives 144, 146; digital 28  
 historical newspapers 14, 25, 39, 53, 58, 69; material 41, 48  
 historical research 20, 41, 58, 161, 166  
*Historic Newspapers in the Digital Age 2*  
 history of journalism and communication research 159  
 Hobbs, Andrew 56  
 hyperlinks 81, 83–85, 116, 118
- impressions 165–166, 170–171  
 information, licensing 98–99  
 information management 121  
 inquiry, historical 158–160, 163, 171  
 intelligence 161–162, 169  
 internships 118, 122  
 interviews 42–43, 46, 99–100, 113, 119–120, 122, 130, 134, 151  
 inverted pyramid model 9–12, 16, 18, 20, 42
- Jacobi, Carina 65  
 Janeiro, Rio de 154  
 Jarvis, Jeff 158  
 Jeffrey, Robin 129–130  
 Jensen, Strandgaard 3  
 Jordan, Michael I. 60  
 journalism 2, 4–7, 9, 11, 14, 39–43, 47, 54, 94–95, 106, 110, 113, 121–122, 135, 158–159; digital archives of 5–7; historical 53–54, 70; historiography 10, 110; research 5, 13, 39–40, 74, 94–95, 114–115; scholars 3, 7–8, 38–39, 54; studies 2–3, 6, 40, 43, 56, 136  
 journalism history: research 9–10, 20, 123; studies 110–112, 116–117, 119, 121, 123; women 110, 112–113, 122–123
- Kvinnsam 110, 113–114, 117–123
- Latent Dirichlet Allocation (LDA) 60, 67  
 Lazer, David 20  
 Lessig, Lawrence 74  
 Lewis, Seth C. 48  
 libraries 3, 74, 94–96, 98, 101–107, 112, 118, 121, 131, 164  
 local news 7, 74, 79–80, 82–83, 85–88
- machine learning 5–6, 38, 40, 45, 47, 172; algorithms 38, 41, 46; approaches, supervised 47  
 Mahrt, Merja 20  
 Manovich, Lev 44  
 Masanès, Julien 76  
 materials 3–5, 7, 18–20, 24, 55–56, 70, 105–106, 111–114, 116, 119, 121–123, 143–145, 148–149, 152, 154–155; source 19, 38, 41, 163  
 Matzner, Deborah 134

- media 5, 12, 15, 19, 25, 29–31, 48, 53, 75, 128–131, 136–137, 158–162, 170–172; archaeology 159, 161–162, 170; coverage 144–145, 148, 153; history 7, 12–13, 25, 34, 120, 158–159, 162, 164, 170; institutions 25
- metadata 13–15, 39, 43, 45, 70
- methodological experiment 158–159, 162, 170
- Mignolo, Walter 146–147
- modal genres 42, 47
- Moretti, Franco 57
- movement of intelligence 159–163, 168–170, 172
- Mussell, James 13, 39
- mystification 29–30
- Napoli, Philip M. 6, 74, 80
- Neiger, Motti 114
- neighborhoods 143–144, 147, 149–151
- Nelson, Robert K. 58, 66
- Nerone, John 11–12
- network analysis, social 83, 88
- networks 83–84, 122, 135, 138, 149, 154, 170
- Newman, David J. 58, 66
- news: applications 6, 75, 96–97, 102, 104–107; apps 96, 99–103, 105–106; archiving 7, 74, 103, 106; business 158–160; media 70, 74, 76, 79, 86, 88, 145, 153, 159–160, 163, 166; media industry 74; organizations 94–96, 98–101, 103, 106–107, 132; sources 10, 13, 80–81
- newspaper archives 6, 38, 53–54, 57, 79; digital 38, 47, 59; historical 39, 45, 53, 56; topic modelling nineteenth-century 6, 53
- newspapers 11–13, 15–16, 18, 24, 31–32, 40–44, 53–56, 59, 62–63, 66, 68–71, 74–76, 78–79, 83–84, 129–130; content 13, 40, 42, 46, 76–79, 88; philosophical 158–160
- newsrooms 94, 96, 98, 100, 102–107, 113, 131
- news stories 10–12, 40, 55, 60, 85, 95, 153; digital 94
- news storytelling 12, 16, 18
- Ng, Andrew Y. 60
- NGOs 145, 149, 154–155
- Nicholson, Bob 6, 13, 53, 64
- Notable New Yorkers of 1896–1899* 167
- online news archives 111, 120, 127
- oral history project 110–111, 113, 121
- Pöttker, Horst 11
- Page, Walter H. 169
- Pandian, Mathias 128
- Peters, John Durham 123
- platforms 76–77, 80, 95, 98, 103, 105, 135, 149, 162
- practices, web archiving 76, 79
- programming languages 77, 94, 98, 101, 163
- qualitative archive analysis 114
- Quandt, Thorsten 59
- questionnaire 94, 99–101, 103, 106–107
- Radford, Jason 20
- remediation 7, 158–160, 162–163, 169, 171; operations of 167, 171; retrograde 162
- resistance 28, 30, 122, 134–135, 137
- Robson, Catherine 166
- Rollins, Frank W. 169
- sampling 4, 12, 45, 56, 59, 70, 81–82
- Scharkow, Michael 20
- scholarship, historical 48, 137
- Shaw, Adrienne 115
- Simon, A. F. 47
- Sims, Norman 169
- Soares, Luis E. 151
- social movements 7, 31, 75, 127–129, 131, 136–137
- social organism 161–162, 165
- solidarity 151, 153–154
- Sonwalkar, Prasun 136
- Soundra, Samuel 128
- sources, digital 9–10, 14
- Star, Susan Leigh 95
- Stecula, Dominik 155
- Tenen, Dennis Yi 4
- Tenenboim-Wenblatt, Keren 114
- Tharakam, Bojja 134
- Theimer, Kate 112
- Thirumal, P. 130
- Thompson, Edward 135
- Thought News* 159–160, 166–169
- Thrall, Trevor 155
- Tontine: What It Is; How It Works* 164
- topic modelling/models 5–6, 40, 43, 47, 53–54, 58–67, 69–71
- Trilling, Damian 43, 58, 136
- tweets 7, 163, 165–168, 170–172
- Twitter 74, 132, 134, 160–163, 165, 169–170, 172
- Twitter bot 159, 161–163, 169–170
- United States 53–55, 62–63, 65, 67, 70, 74, 78–79, 85–86, 96, 100, 106, 160
- Uricchio, William 111
- Vemula, Radhika 134
- Vinton, Lindley 169
- visibility 115–117, 120–121
- visualization 83–85, 96
- Viva Favela 143–145, 149–152
- voices 5, 111, 114–115, 117–119, 121–123, 128, 132, 138, 150, 162, 171; public 122–123

- Wall, Melissa 135  
web archive data 86  
web archiving/archives 6, 74–84, 86–88, 95,  
112, 123; creation of 76, 87; efforts 77–78;  
tools 94, 96  
web data, archived 74, 76, 78, 82  
webpages 75–77, 81–82, 87, 97  
websites 6, 75–77, 80–85, 88, 94–98, 101, 105,  
116, 119, 128, 130–131, 136, 143  
Weischenberg, Siegfried 12  
Welbers, Kasper 65
- Weld, Kirsten 128  
Williams, Raymond 25–26  
women 63, 110–114, 118–123, 130, 134, 150;  
archives 110–111, 114; history 118, 121–122;  
history archives 112–113; journalists 6–7,  
110–111, 113–114, 121, 123  
workflows 60, 94, 96, 103  
YouTube 128, 131–134, 137  
Zamith, Rodrigo 48