Journalism Research
2021, Vol 4 (2) p. 74-97
ISSN 2569-152X
DOI: 10.1453/2569-152X-22021-11525-en

HERBERT VON HALEM VERLAG

Research paper

Marco Braghieri, Tobias Blanke and Jonathan Gray

Journalism aggregators: an analysis of *Longform.org*

How journalism aggregators act as site of datafication and curatorial work

Abstract: **What is the role and significance of digital long-form content aggregators in contemporary journalism? This article[1] contends that they are an important, emerging object of study in journalism research and provides a digital methods analysis and theoretical engagement with *Longform.org*, one of the most prominent long-form content aggregators on the web. We propose that *Longform.org* can be understood as leveraging the datafication of news content in order to valorize the long tail of archived material. Drawing on scraped data from the archive, we undertake an in-depth analysis into the practices of long-form aggregators. While *Longform.org* exhibits a degree of curatorial diversity, legacy news media outlets tend to be featured more frequently. Accessibility of news media archives is one of the most relevant factors for being featured by *Longform.org*. Our analysis demonstrates the relevant role of smaller digital-only publications, which provide a unique mix of sources. Through a network analysis of scraped tags we explore the composition of themes, including personal, world-political, celebrity, technological and cultural concerns. The data and curatorial practices of such long-form aggregators may be understood as an area of contemporary news work that conditions which past perspectives are more readily available, experienceable and programmable on the web.**

1 A similar version of this article has appeared as a book chapter in: Marco Braghieri (2021): *Yesterday's News. The future of long-form journalism and archives*. Oxford et al.: Peter Lang.

Introduction

While in contemporary digital news consumption, »media stories monopolize the attention for a week or so and then are instantly forgotten« (Fisher 2009: 59), news media digital archives are the recipients of a relevant interest over the World Wide Web (cf. Elliott 2012). Such archives are intended here as the result of »journalistic publications, productions or related content [...] stored and made available in digital form« (Bødker 2018: 1114). While long-form journalism can also be defined as slow journalism (cf. Le Masurier 2015) and literary journalism, »true life stories that can be read like a novel or a short story« (Hartsock 2000: ix), within contemporary digital news production and consumption, long-form journalism departs from the accelerated news cycle (cf. Le Masurier 2016); values writing standards and research (cf. Le Masurier 2015); searches for originality (cf. Belt/South 2016); and places a range of multi-modal digital media production techniques at »the heart of its narrative structure« (Hiippala 2017: 421).

Thus, long-form journalism must be framed as part of digital journalism, »the transforming social practice of selecting, interpreting, editing and distributing factual information of perceived public interest to various kinds of audiences in specific, but changing genres and formats. As such, digital journalism both shapes and is shaped by new technologies and platforms, and it is marked by an increasingly symbiotic relationship with the audiences« (Steensen et al. 2019: 338). Thus, it is important to stress how, as underlined by Seaton (2016), the contemporary architecture of communications is defined by two polarities: »there is an overwhelming abundance of information and communications that are multifaceted and shared multilaterally and multinationally« and yet »narrow ›silos‹ of information and opinion may develop« (Seaton 2016: 808). The scenario described by Seaton (2016), is also defined by the economic difficulties faced by legacy media, as stressed by Franklin: »characterised by falling audiences, readerships and advertising revenues« (Franklin 2014: 482). The media industry has reacted in part investing in »new platforms and consciously diversifying their product portfolios«, yet without a certain outcome: »it is not clear whether media corporations will reap the kind of profits they envisage or news consumers will adopt their new products with the readiness they expect and forecast« (Chyi/Chadha 2012: 432). While mobile usage has gathered traction, as underlined by Nel and Westlund (2012), the media industry has once again found itself at a crossroads between re-imagining its approach or choose a more passive stance. The latter would lead to »independent developers [...] leaping at the opportunity to create apps that harvest the rich content newspapers make available freely on the Web« (Nel/Westlund 2012: 751).

This forecast finds an echo in the long-tail model of the web economy by Anderson (2009), as distribution in the digital contemporary is performed more

efficiently by aggregators rather than producers. Anderson defines the long-tail as a model defined by »infinite choice. Abundant, cheap distribution means abundant, cheap, and unlimited variety« (Anderson 2009: 180). However, the sales cost has to be as low as possible otherwise market entities become ›entrenched industries‹ (Anderson 2009: 185). Anderson indicates the news media as an example of entrenched industry and underlines the value-generation capacity of intermediaries, or »aggregators« (Anderson 2009: 88). This article builds on this idea by Anderson and investigates how long-form journalism is distributed through aggregators, especially by dedicated entities such as *Longform.org*.

Longform.org has been an object of significant attention, either presenting reader data (cf. Boynton 2013), or describing its inception and nature (cf. Albalad Aiguabella 2015) or in the description of the digital news ecosystem analyzing long-form journalism (cf. Dowling/Vogan 2015; Longhi/Winques 2015) and contemporary readership (cf. Jacobson et al. 2018). This article attempts to place *Longform.org's* activity in relation with digital news outlets' archives and with one of its most relevant types of content, long-form journalism, framing this analysis within Anderson's long-tail theory (cf. Anderson 2009). Moreover, it provides, through digital methods (cf. Rogers 2013; Venturini, Bounegru et al. 2018) – repurposing »methods of the medium« such as scraping and hyperlink analysis – a fresh engagement with the role and operations of news aggregators. We work with a dataset of over a thousand posts on *Longform.org* from 2016.

Background: News Outlet Archives and Aggregators as distributors of long-form journalism

This section is dedicated to creating a framework on how news outlet archives and aggregators' activity can be framed as a distribution practice. It is important to underline that news media outlet digital archives are capable of attracting relevant interest on the World Wide Web. According to the former *Guardian* readers' editor, Chris Elliott, »*The Guardian's* digital archive holds more than 1m articles [...] And it is very popular. Nearly 40% of content viewed on the website is more than 48 hours old« (Elliott 2012). Moreover, a study on long-form journalism by Smith, Connor and Stanton (2015) established how within their corpus of 5.2 million long-form journalism stories, relevance over time was a defining factor as »longform articles tend to maintain external links, a proxy for interest, longer than typical news articles« (Smith et al. 2015: 2115).

However, this retrieval process poses distinct challenges to content management systems which within newsrooms are used for a number of activities ranging from content creation to editing, publishing and distribution (cf. Barker 2016). Thus, content retrieval in general, and long-form journalism retrieval from news

Screenshot Longform.org (18 June 2021)

LONGFORM

Home
Best Articles
Podcast
Sections
Collections
Reprints

Search

Random Article

Lists
Best of 2020
Best of 2019
Best of 2018
Best of 2017
Best of 2016
Best of 2015
Best of 2014
Best of 2013
Best of 2012

Archive
Publications
Writers
Tags

More
Newsletter
Suggest Article
About
Contact
Advertise

Get our Newsletter
Great articles, every Saturday.
Your email address
Subscribe

Read Later
Choose service...

Today, June 18

SCIENCE HEALTH

Is There Something Wrong With the Air in South Portland, Maine?

Residents have lived near more than 100 massive petroleum storage tanks for decades, never really knowing if they're breathing in dangerous chemicals. Now they're fighting to find out.

KATHRYN MILES BOSTON GLOBE MAGAZINE JUN 2021 15MIN

Yesterday, June 17

My Father Vanished When I Was 7. The Mystery Made Me Who I Am.

My dad was a riddle to me, even more so after he disappeared. For a long time, who he was—and by extension who I was—seemed to be a puzzle I would never solve.

NICHOLAS CASEY NEW YORK TIMES MAGAZINE JUN 2021 35MIN

SCIENCE

The Lab Leak Theory Doesn't Hold Up

The rush to find a conspiracy around the COVID-19 pandemic's origins is driven by narrative, not evidence.

JUSTIN LING FOREIGN POLICY JUN 2021 20MIN

HISTORY

If We Can Soar: What Birmingham Roller Pigeons Offer the Men of South Central

In 1970 South Central, pigeon fancying was serious business. But there's a deeper story behind why these Black Angelenos are entering their fifth and sixth decade raising Birmingham Roller pigeons.

SHANNA B. TIAYON PIPE WRENCH JUN 2021 30MIN

outlet archives more specifically, poses a difficult challenge within the usage of content management systems. Long-form journalism is a type of content that maintains its capacity to attract readers over time and, as such, is more likely to be made available by news outlets. Hence, we can define long-form journalism as one of the factors promoting interest in new media outlets' archives. More broadly, within the digital contemporary, single news stories are organized in digital news archives that operate as distribution tools. Hence, we shall now focus on how news outlet archives can be envisioned as content distributors.

Anderson describes the long tail as a model based on »infinite choice« (Anderson, 2009: 180). The long tail has established itself as one of the leading production and distribution models within the digital economy and society. According to this model, distribution has become cheaper, and variety has been amplified, with audiences tending »to distribute as widely as the choice« (Anderson 2009: 180). According to Anderson (2009), this new model has been embraced more efficiently by intermediaries rather than traditional producers. Anderson (2009) mentions the media industry as one of the main examples of this dynamic: aggregators are performing more efficiently in distributing content if compared to traditional news media outlets. Moreover, there are other relevant factors in the long tail, such as the democratization of production and distribution tools and the connection through filtering between supply and demand (cf. Huang/Wang 2014).

Within the digital contemporary, long-form journalism and digital news media outlet archives are yet to be fully datafied (Mayer-Schönberger/Cukier 2013). Mayer-Schönberger and Cukier define »datafication« as the process of organizing a phenomenon »in a quantified format so it can be tabulated and analysed« (Mayer-Schönberger/Cukier 2013: 78). However, while news media outlet archives in the digital contemporary have a growing digital presence, stemming from the digitisation of physical archives, harmonization of digital archives and digitally native archiving practices, its content is yet to be datafied. As Blanke and Prescott underline the datafication process is ›different from the process of producing digital surrogate based on digitising originally analogue content by [for example] transferring a microfilm of a book to digital form or making an MP3 version of a taped interview‹ (Blanke/Prescott 2016: 192). Hence, datafication is based on the principle that the process outcome can be transformed in a quantifiable format for it to be exploited in different manners.

However, datafication is not a neutral process, as some aspects recall the issues raised by what Derrida defines as the »de-paperization« process (Derrida 2005). While Derrida (2005) identified its potential benefits, he also underlined the issues it raises, such as »invisible hegemonies and appropriations« (Derrida 2005: 55ff). In this regard, an example of the non-neutrality of the datafication process is provided by Mayer-Schönberger and Cukier's praise of the Google Books project as an example of successful datafication (cf. Mayer-Schönberger/Cukier

2013). In 2004, Google began scanning books, gradually building a digital library and, by 2015, the Google Books project had scanned »more than 25 million volumes [...] including texts in 400 languages from more than 100 countries« (Heyman 2015). This process, aside from creating an immense digital library, created economic value for Google (Pybus et al. 2015). Hence, while this datafication process can be defined successful for Google, leading to the creation of an asset with sizable economic value and multiple future applications, it has likewise produced an »invisible hegemonies and appropriations« (Derrida 2005: 55ff.) as Derrida warned while describing the potential issues with the »de-paperization« process.

Before beginning our analysis of a single long-form journalism aggregator across one year of activity, it is useful to remind ourselves of the investigations by Smith, Connor and Stanton (2015) regarding long-form journalism production. The authors underline how overall long-form journalism production in the digital contemporary is increasing, yet it is doing so following a specific trajectory. As the study's authors emphasize, there is an increasing number of news media outlets producing long-form journalism and, among these, there is an increasingly strong presence of digitally native news media outlets (cf. Smith et al. 2015).

Hence, these findings (Smith et al. 2015) describe the digital contemporary as an environment where long-form journalism can be found in numerous news media outlets, but only a relatively small number of those have the economic and organizational strength to produce long-form journalism on a continuous basis, as also underlined by Bruns, Highfield and Lind (2012). »Few journalistic organizations can afford to engage in much long-form, resource-intensive, investigative journalism« (Bruns et al. 2012: 2). The increase in long-form journalism production in the digital contemporary is the result of a large number of news media outlets producing small quantities of long-form journalism stories, while increasing at a slower pace if compared to standard news production, this specific type of journalistic production is capable of remaining relevant for more extended periods of time (Smith et al. 2015). Thus, within the scenario described by Smith, Connor and Stanton (2015), focusing long-form journalism through the analysis of an aggregator such as *Longform.org*, is an effective way of investigating the production and distribution of this specific form of journalism.

An Example of Long-form Journalism Aggregation and Curation: *Longform.org*

As we have seen in the previous section, according to the long-tail model, third-party content aggregators generally perform more efficiently distribution tasks, if compared to traditional entities and the news industry is an example of this phenomenon (cf. Anderson 2009). In this section, we shall focus on the data-dri-

ven analysis of a single long-form journalism aggregator, *Longform.org*.

Longform.org (cf. Longform.org 2010) was founded in 2010 and begun by recommending recently published and digitally archived news items which were over 2,000 words long and already available on the World Wide Web. The difference between *Longform.org* and similar aggregators is the objective of its activity, as underlined Shapiro, Hiatt and Hoyt (2015). *Longform.org* »doesn't hog the traffic; it simply pushes readers on over to the host site« (Shapiro et al. 2015: 175).

Longform.org's first online version dates back to 2010. It was released in parallel with the launch of Apple's iPad, by two individuals, Max Linsky and Aaron Lammer. Since 2010, the aggregator has broadened its offer. In 2012, it added a fiction section, begun an intense podcast production and developed its first iPad application, which was priced at 5$ (USD) and, as 2014, sold circa 60.000 copies (cf. Bercovici 2014). Moreover, according to an article published in *New York* magazine in 2014, the podcast service had reached 50.000 listeners (cf. Kachka 2014). In September of the same year, *Longform.org* released its first iPhone application, introducing the possibility for readers to build lists of specific writers to follow across different news media outlets and developed an algorithm regulating which articles were to be featured within the application (cf. Mullin 2014). However, as of April 2017, *Longform.org*'s application is not available anymore on both iOS or Android platforms (cf. Longform.org 2017b). Founders Max Linsky and Aaron Lammer explained that this is due to the rejection by Apple of their newly developed version of the application (cf. Longform.org 2017b).

While Linsky and Lammer underline how the tools that allow readers to access long-form content have significantly developed over the years, they also stress that – through *Longform.org* – they »have sent over 100 million outbound links to publishers since 2012« (Longform.org 2017b). Moreover, it is relevant to underline how *Longform.org* perceives itself as »closer to a technology company than a communications medium«[2] (Albalad Aiguabella 2015: 18).

Further insight on *Longform.org* is provided by Robert S. Boynton. According to Boynton's data, »Longform's demographic is the envy of any advertiser: young (fifty percent of the readers are under 34), mobile (thirty percent read primarily on phones or tablets), and well educated (forty-two percent have attended graduate school)« (Boynton 2013: 130). Moreover, he underlines how »the best narrative non-fiction – unlike basically every other content type on the web – doesn't lose appeal as it ages [...] Longform's readers are ten percent more likely to read an older story than a new one. The publication date carries almost no weight. Readers care more about an article's subject than whether it is new« (Boynton 2013: 130ff.). Hence, according to Boynton's data long-form journalism and archives

2 The original article is in Spanish: ›*Longform.org* – apunta su propietario – se aproxima más a una empresa de tecnología que a un medio de comunicación‹

in the digital contemporary share not only relevance over time but also possible aggregation and curation practices which, if performed with a user-centric approach such as the one implemented by *Longform.org*, can provide a steady flow of readers to content which news media outlets host on their archives but which is otherwise inactive.

For news media outlets, the activity performed by aggregators such as *Longform.org* demonstrates that, within their digital archives, there are pools of resources, specifically long-form journalism stories, which are under-used as they are not aggregated and curated in order to enhance readership. While this aspect is relevant for legacy publications, which own a great deal of archived content, it is also significantly relevant for news media outlets which have a shorter lifespan, as confirmed by readership data originated by *Longform.org*, as described by Boynton: »A well-known publication name doesn't move the needle much at all [...] unknown publications often do better than brand names because readers are intrigued to see something new« (Boynton 2013: 131).

Having gathered insight on *Longform.org's* creation and the data resulting from its activity, we shall now confront them with a specific dataset we have extracted from *Longform.org's* activity in 2016. Thus, we shall assess if the critical factors identified in this section, such as the relevance over time of long-form journalism and the role of archives as resource pools, are coherent with the data analysis in the following section.

Longform.org's Activity – a Quantitative Analysis

We shall now produce a quantitative analysis of all the entries produced by *Longform.org* in 2016. The generation of our dataset began by collecting all *Longform.org* entries from 1 January 2016 to 31 December 2016. *Longform.org*'s website provides a page by page navigation that goes back to 1 April 2010 and, while the website's design has changed since its inception, it has maintained its organization around a central column which features a feed of articles.

We obtained the necessary data for all *Longform.org*'s 2016 via the World Wide Web, through a process named web scraping which can be defined as ›the practice of gathering data through any mean other than a program interacting with an API (or, obviously through a human using a web browser)« (Mitchell 2015: viii). Moreover, »scraping is not only a technique but equally involves a particular way of dealing with information and knowledge: it is also an analytic practice«. (Marres/Weltevrede 2013: 317). As such, web scraping is being used within the framework of digital methods, intended as »techniques for the ongoing research on the affordances of online media« (Venturini/Bounegru, et al. 2018: 4), deployed to harvest »information made available by Internet platforms« (Venturini/Bou-

negru, et al. 2018: 2). In order to perform our web scraping operation, we used a browser extension for Google Chrome, named Data Miner »that assists you in extracting data that you see in your browser and save into an Excel spreadsheet file« (Data Miner 2016).

To obtain the information we needed from *Longform.org*, we had to develop a series of »extraction instructions that Data Miner uses to extract data from websites« (Data Miner 2016), which are named recipes. We were able to scrape 50 pages out of *Longform.org*'s website. The data was provided in a comma-separated values file (CSV), which we then fed to OpenRefine, an Interactive Data Transformation Tool (cf. Verborgh/De Wilde 2013). Through OpenRefine, we performed a data profiling and data cleaning processes. Data profiling was implemented to »discover the true structure, content and quality« (Olson 2003: 119) of the scraped data. The data cleaning process was implemented in order to correct possible errors in our data ›in a semi-automated way‹ (Verborgh/De Wilde 2013: 6). Hence, we shall now analyze the resulting data set which is the outcome of our web scraping, data profiling and data cleaning processes.

In total, we scraped 1.225 posts from 1 January 2016 to 31 December 2016[3]. Typically, *Longform.org* elaborates posts which comprise a link to a single long-form journalism story, a summary and information on the author, news media outlet and date in which the long-form story was first published. In 2016 it published 1.074 single-story entries. However, alongside this primary type of entry, *Longform.org* has developed, over the years, other types of entries. The first is the ›Longform guide‹ entry, which typically groups together long-form journalism stories from different news media outlets which focus on the same subject (30 in 2016). *Longform.org* also publishes entries dedicated to a single author, which has already been featured multiple times in the websites single story entries (15 in 2016). Besides the ›Longform Guide‹ entries and the entries dedicated to single authors, there are weekly entries dedicated to fiction writing (51 in 2016). Aside from its long-form aggregation and curation activity, *Longform.org* has developed a significant original multimedia production, through a podcasting series (55 in 2016).

As we can see, the vast majority of entries regarded single story entries, which are the ones we shall analyze in-depth. We will provide an assessment of the news media outlets selected by *Longform.org* to be featured in this type of entries, and we shall also focus on the publishing dates in which each long-form journalism story was first published. As *Longform.org* is a curation service which we have framed as an aggregator following the long-tail model (Anderson, 2009b), it is relevant to examine its choices in detail, as its activity revolves specifically around long-form journalism and news media outlets digital archives in the digital contemporary.

3 The dataset which is used in this article is publicly available and has been uploaded onto the Open Science Framework website at https://osf.io/8myj5/

We shall concentrate our analysis on the 1.074 single story entries published by *Longform.org* in 2016. As the first publishing date for each long-form journalism story is provided within the entry, it is possible to establish how *Longform.org* has distributed its choice within news media outlets and different production eras. The 1.074 single story entries on *Longform.org* in 2016 are drawn from an extensive time frame, as the oldest long-form journalism story that was featured in a single-story entry was first published in 1877. Year-wise, the most relevant group is the one which comprises stories published in 2016, the same year which we focused our analysis on, as single-story entries based on long-form journalism stories first published in 2016 were *77%* of the total.

Figure 1

Single Story Entries on *Longform.org* in 2016 divided per first publishing date

As we can see in Figure 1, the timeframe from which long-form journalism stories were chosen to be featured in single story entries is very ample, approximately in the period ranging from 2009 onwards, choices tend to become more

frequent. *Longform.org* aggregates and curates solely content within news outlet websites or archives, which do not implement a »radical« paywall (Brock 2013: 155). This pre-condition does appear to affect the total pool of news media outlets this specific aggregator uses.

Moving on from the overall distribution of first publishing dates of *Longform.org* single story entries is clear, we focus next on those entries which have a first publishing date that precedes 2016, the year of *Longform.org's* activity on which our analysis is focused on. As we have seen in Figure 1, they comprise 23% of the total single-story entries, but – in order to gain better insight on their distribution – we have divided all pre-2016 entries in decades, grouping together all entries referring to long-form journalism stories published before 1960.

Figure 2

All *Longform.org's* 2016 single story entries with a first publication date prior to 2016

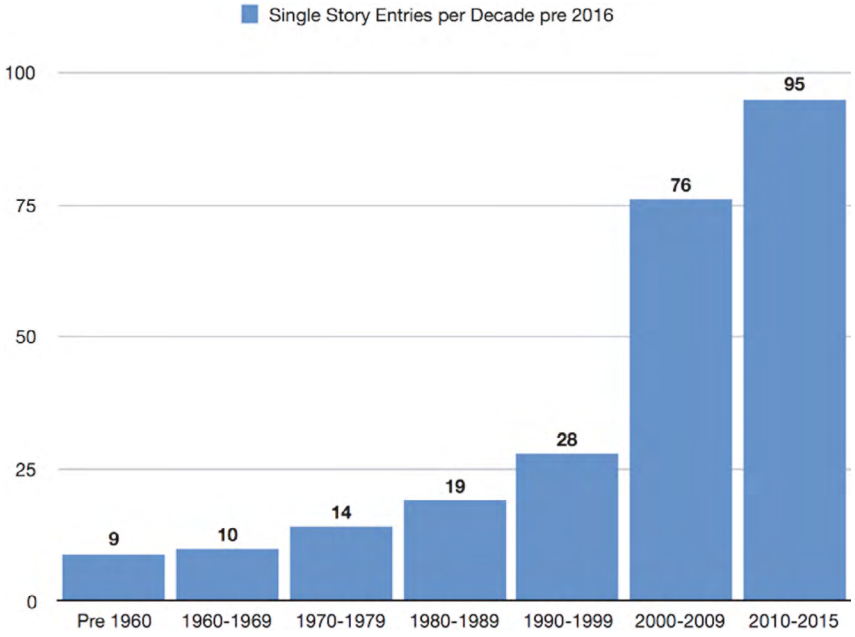


Figure 2 shows how single-story entries which revolve around long-form journalism stories with a first publishing date that pre-dates 2016 increase gradually every decade. However, the increase in entries dated between 1990-1999 and the

following decade is a record 171 per cent. The newsroom digitisation process has had a significant impact on news media outlet digital archives, broadening their development and fruition, which seems to be confirmed by the number of stories sourced from the 2000-2009 decade. While digital news media outlet archives which originate from physical copies are labor-intensive to create, the newsroom digitisation process has brought broader access to news archives. As we can see in Figure 3, news media outlets which have significantly developed their digital news archives tend to be more represented.

Figure 3 shows that slightly more than four out of ten long-form journalism stories selected in single story entries by *Longform.org* were published on outlets which are overall featured ten times or less in the 1,074 entries. The total number of news media outlets featured in single story entries by *Longform.org* in 2016 is 221 and, out of these, only 25 news media outlets are featured ten or more times. Moreover, among the 196 news media outlets which have been chosen for ten or fewer entries, the most relevant group comprises outlets selected once or twice. Out of the 196 news outlets featured less than ten times, the most relevant group are the news media outlets that have been chosen just once, comprising 108 news media outlets.

While the majority of *Longform.org's* selection revolves around a selected number of news media outlets, variety in news media outlet selection is a relevant factor among *Longform.org's* selection choices. This diversity is achieved not just by generally widening the number of news media outlets the aggregator sources its stories from, but by specifically choosing news media outlets which are featured fewer times. This selection activity seems to indicate that one of the major focuses in aggregator activity is variety in outlet selection. However, among news media outlets featured in single story entries in 2016, there has been a particular focus on The New Yorker, featured in 95 entries.

To assess the consistency of the top news media outlets featured in single story entries in 2016, we compared our dataset with the overall number of times a news media outlet has been featured on *Longform.org*. However, while this data is available directly on a specific page of *Longform.org's* website (Longform.org 2017a) and is relative to 22 May 2017, the methodology with which *Longform.org* has derived its data is not specified, as the page simply displays the number of news outlets which are featured in more than a set number of posts.

Figure 4 confirms the consistency between the data which we scraped from 2016 single story entries and the overall data provided by *Longform.org*. We can observe how there is a small number of outlets that have not been chosen as frequently in 2016, while being featured extensively in the past, such as *The New Republic*, *Slate*, *The New York Review of Books*, *WIRED*, and *Rolling Stone*. However, it is possible to assume that the top 25 contributors in 2016 for single story entries are representative of the overall news media outlet selection in *Longform.org*.

Figure 3

News media outlets featured in *Longform.org* single story entries in 2016

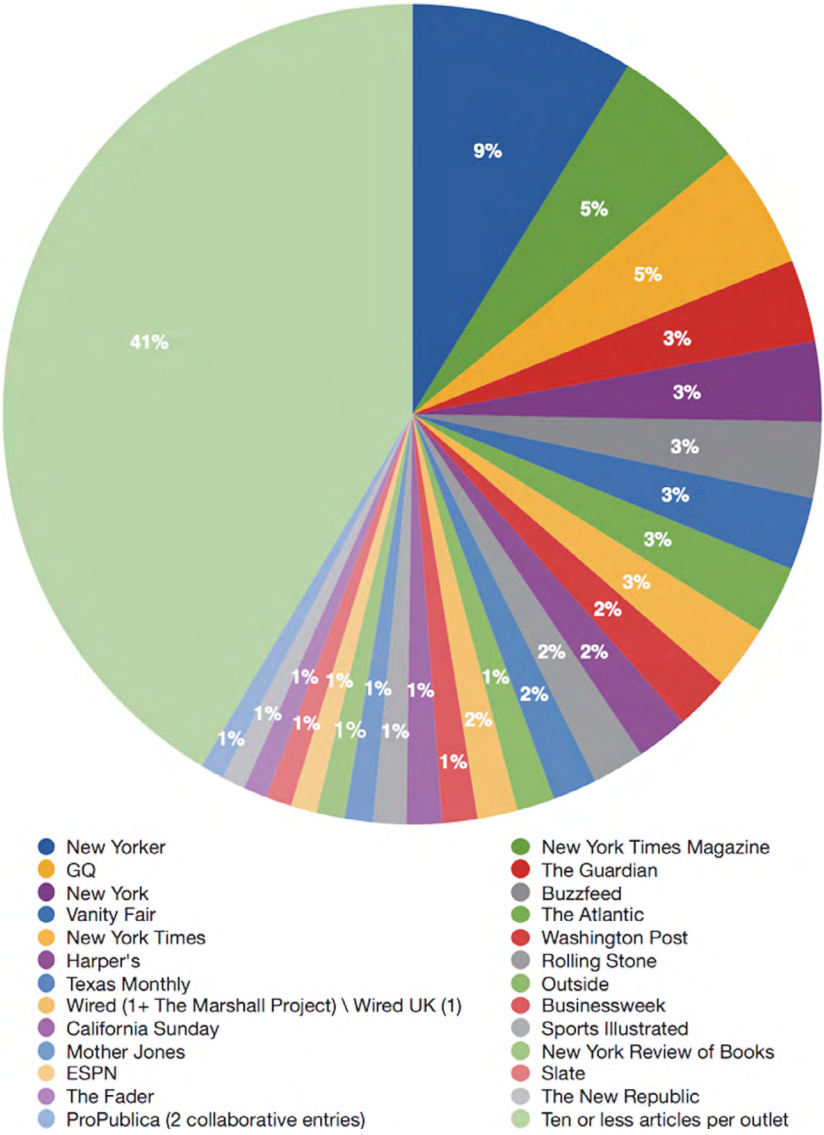
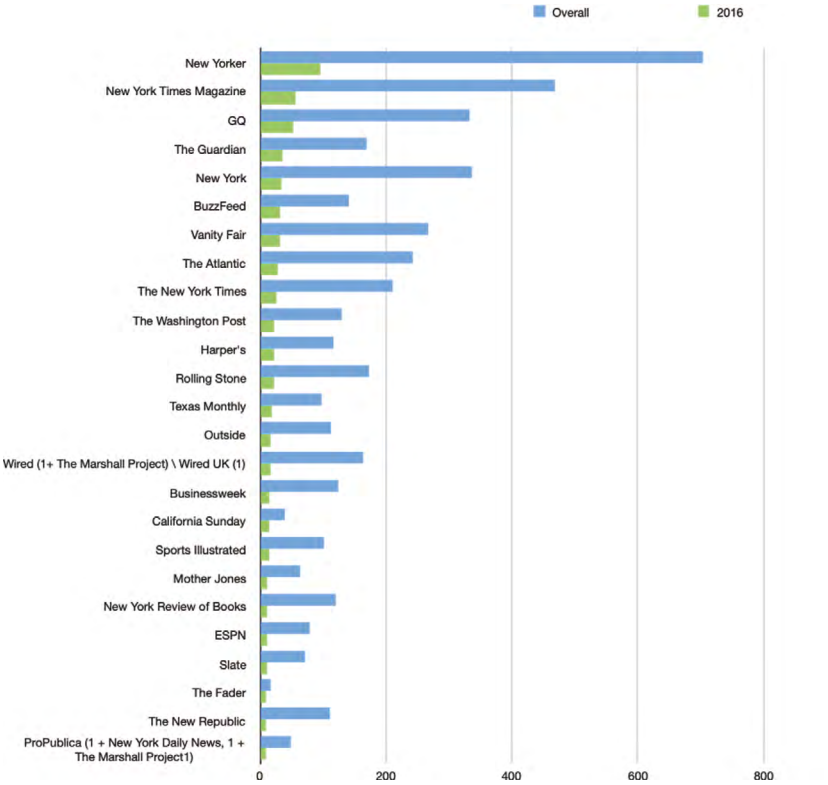


Figure 4

Top 25 news media outlets per single story entries in 2016 and the respective data regarding overall production sourced from *Longform.org's* website.



We can also see how, alongside outlet variety, the other factor at play in news media outlet selection operated by *Longform.org* is legacy. As outlet presence seems to be consistent in both datasets, the overall orientation in *Longform.org*'s choices for its main contributors is decisively aimed at major news media outlets which have been active for an extended period of time, as among the top overall contributors we can find *New Yorker* magazine, *The New York Times Magazine*, *New York* magazine, *GQ*, and *Vanity Fair* magazine.

To further assess news media outlet relevance, we shall focus on news media outlets which have had at least two long-form journalism stories featured in *Longform.org* 2016 single story entries. The number of news media outlets which have been featured at least twice in *Longform.org* 2016 single story entries is 111,

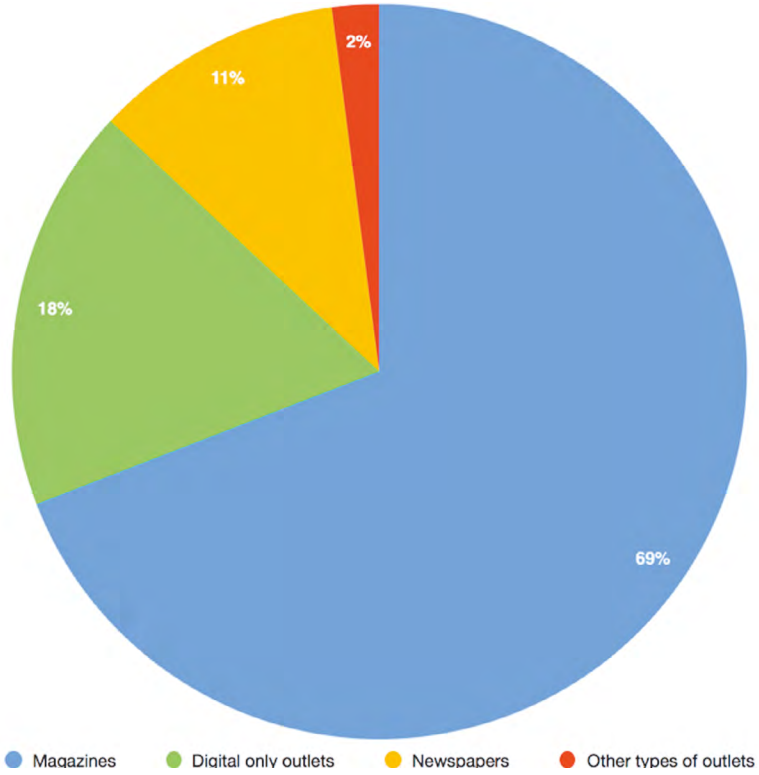
and they are responsible for 965 entries out of the total 1.074. As we have seen in Figure 4, news media outlets most frequently chosen in 2016 are consistent with the overall choices made by *Longform.org* since its inception in 2010. An analysis of the type of news media outlet responsible for these 965 entries could then provide more general insights about which type of news media outlets are more relevant among *Longform.org's* overall choices.

Out of the total 111 outlets which are featured at least twice in single story entries in 2016, 68 are magazines which retain a paper edition, 32 are digital-only news media outlets, 8 are daily newspapers and 3 are websites developed by media companies which focus primarily on another type of medium, such as television or radio. The overall number of magazines which retain a paper edition is highly relevant, both if compared to other types of news media outlets featured in 2016 and to the top overall contributors to *Longform.org*. However, there is a significant number of digital-only publications (32 in total) which are the second largest group among news media outlet types. The number of digital-only publications is moreover relevant as this group is four times larger than daily newspapers, which are only eight, namely the UK's *Guardian*, the United States' *The New York Times*, *The Washington Post*, *The Boston Globe*, *The Los Angeles Times* and *The Tampa Bay Time*s as well as the Norwegian newspaper *Dagbladet*. Alongside magazines, digital-only publications and newspapers, there are three websites developed by media outlets which focus primarily on television or radio, which are *ESPN*, *MTV*, *NPR* and *Fusion*.

Figure 5 displays the percentage of each news media outlet category among the overall single-story entries in 2016. Magazines are the most relevant type of news media outlets, accounting for 69% of all single stories drawn from a pool of 68 different outlets, which were 61% of the total outlets. Digital-only news media outlets account for 18% of all single stories featured but represented 29% of the total news media outlets selected by *Longform.org*. Moreover, while daily newspapers account for 11% of the single-story entries, they represent just 7% of the overall news media outlets. A similar result can be found among other types of outlets which account for 4% of the number of single-story entries but represented just 2% of the overall news media outlets.

The percentage difference between news media outlet type and the number of articles selected from each group shows that digital news media outlets are used as a source for a smaller number of articles. Thus, news media outlet variety is achieved by *Longform.org* mostly by choosing long-form journalism stories published by digital news media outlets. In comparison, magazine and daily newspaper production are both over-represented in relation to the total number of news media outlets in the two categories. Thus, *Longform.org* tends to privilege legacy news media outlets, as we saw in Figure 3. There is a further factor of influence in *Longform.org's* choices which can be clearly identified, as all but eleven outlets

Figure 5
Number of entries per news media outlet type in 2016 on *Longform.org*

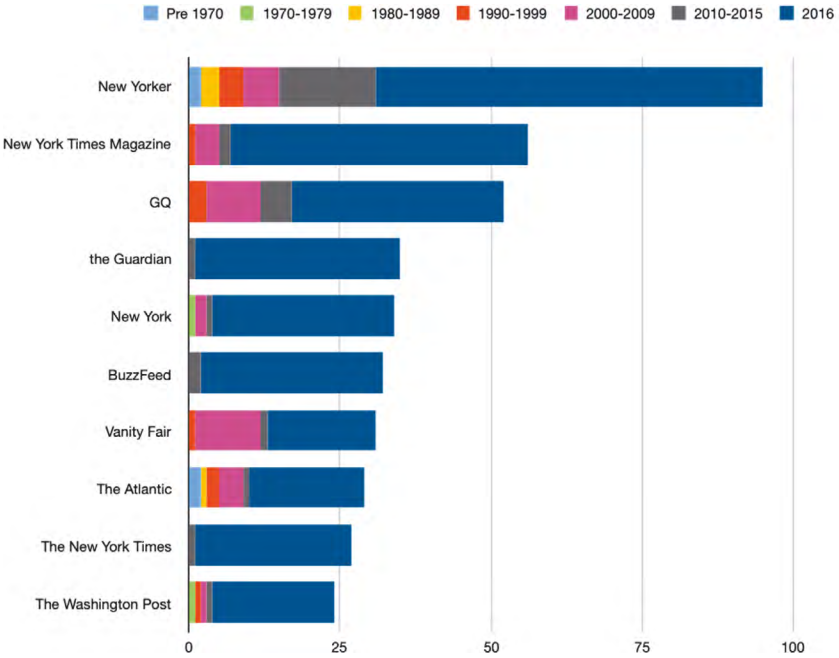


are based in the United States. These news media outlets are based in the United Kingdom, Germany, Norway, Canada and Australia and are *The Guardian*, *Der Spiegel*, *London Review of Books*, *Dagbladet*, *The Globe and Mail*, *Canadian Business*, the *BBC*, *The Economist*, *The Sidney Morning Herald*, *The Toronto* Star and *Toronto Life*. Among these news media outlets, *The Guardian* is the only one with a significant impact in terms of selection, as it accounts for 35 single-story entries in 2016. Overall, the number of single-story entries sourced from non-US based news media outlets is 56 out of 1074.

To further examine how specific news media outlets are selected by *Longform.org*, we shall now focus on the first publication dates of long-form journalism stories sourced from the top ten contributors in 2016.

Figure 6

The top ten news media outlets in 2016 and the first publishing dates of their long-form stories divided per decade and in the year 2016



As we saw in Figure 3, single story entries in 2016 were sourced from a wide variety of news media outlets. Out of the top 25 news media outlets which were featured more than ten times in 2016, we shall focus on the top ten. In order to better understand how single-story entries from these news media outlets have been selected we have tracked the publication date for each of the entries by *New Yorker*, *The New York Times Magazine*, *GQ*, *The Guardian*, *New York* magazine, *BuzzFeed*, *Vanity Fair*, *The Atlantic*, *The New York Times* and *The Washington Post*. We have divided first publication dates into seven different time frames, one devoted to entries first published before 1970, followed by five decade by decade time frames, devoting the last time frame to entries first published in 2016. As we can see in Figure 6, the latter is by far the most represented category. However, only two news media outlets have entries which were first published in six different time frames, *New Yorker* and *The Atlantic*. Both are legacy news media outlets, as they were founded in 1925 and 1857 respectively. Moreover, we can see how *The New York Times Magazine*, *GQ* and *Vanity Fair* all have a similar pattern, with entries from the same

four eras [1990-1999; 2000-2009; 2010-2015 and 2016]. The only exclusively digital outlet featured among the top ten is *BuzzFeed*, yet its pattern is similar to the one developed by entries from *The Guardian* and *The New York Times*, which feature only single-story entries first published between the 2010-2015 period and in 2016. *The Washington Pos*t highlights a different pattern, with entries drawn from five different time frames. Hence, we can conclude that legacy magazines entries tend to be drawn from a more composite pool of publishing eras, with a more pronounced focus on entries which were first published before 2016.

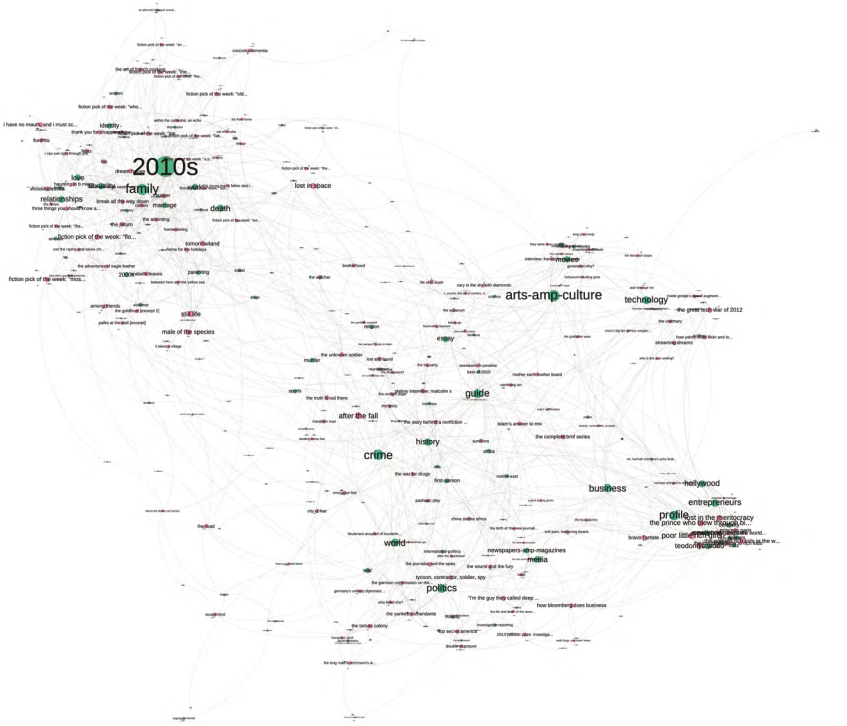
Finally we explored the relations between articles and their associated tags by creating a network diagram (see Figure 9) using scraped data that was visualized using the Gephi software and the Force Atlas 2 spatialisation algorithm, such that articles sharing similar tags were clustered closer together (Jacomy et al., 2014). Using this network as a device to visually explore associations (Venturini, Jacomy, et al., 2018) in aggregated content through *Longform.org's* tagging practices, we can discern five main clusters of articles. Firstly, perhaps the most prominent cluster in the top left, of which the largest tag is »2010s« (associated with three quarters of all of the articles) contains themes such as »love«, »relationships«, »sexuality«, »identity«, »women«, »family«, »marriage«, »parenting«, »work«, »death«, suggesting the thematization of the personal and narratives about experiences of everyday life. Secondly, a tighter cluster in the bottom right is concerned with »hollywood«, »entrepreneurs«, »celebrity«, »profile« as well as »dictators« and »cults«. Thirdly, a region to the top right concerns »technology«, »dot-coms« and »gadgets«. Fourthly, an adjacent area focusing on »movies«, »film« and »arts & culture«. Fifthly and finally, a more diffuse region towards the center-right contains articles associated with »crime«, »history«, »world«, »politics«, »war«, »international politics«, »germany«, »cia«, »cuba«, »white-house«, »afghanistan-war«, indicating an enduring interest with dramatic events on the global stage.

Tags representing journalistic genres across the network include: Crime (13%), Arts and Culture (8%), Essays (7%), Profiles (7%), Politics (7%), First Person (6%), Business (6%), Sports (6%), Technology (5%), Science (5%). From this brief analysis we can see how the articles tagged on *Longform.org* indicate the resonance of personal, dramatic world-political, celebrity, technological and cultural themes in long-form news archives.

Conclusion

In this article, we focused on the *Longform.org* aggregator, providing data on its aggregation and curation choices. Assessing how these third parties perform their activity, allowed us to identify a specific set of practices, such as news media outlet variety and a balance between more recent and older long-form journalism stories.

Figure 7
Bipartite network of articles and their associated tags generated using Gephi visualization software



Among the news media outlets selected by a third-party aggregator such as *Longform.org*, we have found how legacy news media outlets tend to be featured more frequently as they provide a mature, more diverse pool of long-form journalism stories. Thus, within the aggregation of long-form journalism, news media outlet archives assume a relevant role in allowing older content to be redistributed.

Moreover, *Longform.org* content aggregation and curation activities adhere to the intermediary role as theorized by Anderson (2009) and his ideas of long-tailed distribution and consumption on the web. Anderson's focus looked at the efficiency of the distribution process, rather than towards the ownership of a specific product (cf. Anderson 2009) – in our case in exam, long-form journalism. Incumbent entities usually outperform traditional industries by using new dis-

tribution practices at scale as in the case of platform services such as social networks which – among other activities – perform an intermediary function between producers and crowds. *Longform.org's* case, in this sense, is highly relevant as it did not begin as a scale operation and, as the two founders remarked, »the audience that came just kept getting bigger and bigger without us doing much« (McQuade, 2015). Hence, we can assume that, within the digital contemporary, aggregation and curation are decisive factors in the growth of intermediaries, whereas production and ownership's role is diminishing in importance. The success of *Longform.org* as a long-form journalism aggregator in the digital contemporary demonstrates that this type of entities has successfully attracted readers and attention, whereas news media outlets have struggled to develop effective intermediation practices which regard their own content.

There are multiple directions that aggregation and curation activities could take driven by text and data mining processes, especially if the datafication process of news media outlets digital archives will progress in the future. Studies based on person-centric mining (cf. Coll Ardanuy et al. 2016) and based on historical geospatial data extraction (cf. Yzaguirre et al. 2016) both indicate that there seems to be a fertile space for new types of user-centric aggregation and curation services originating from news media outlets' digital archives, once they are datafied. These new types of curation and aggregation seem to be tailored for news media outlets looking to develop aggregation and curation services among their archived production.

Attending to the practices of aggregators such as *Longform.org* may help us to understand how news outlets are organizing online encounters with archives and reshaping how audiences relate to stories of the past – including through recontextualization, recombination, re-valuation and circulation on digital platforms and infrastructures. The data and curatorial practices of such aggregators may be understood as an area of contemporary news work that conditions which past perspectives are more readily available, experienceable and programmable on the web.

About the authors

Dr. Marco Braghieri is a research assistant in Social Big Data at the Department of Digital Humanities, King's College London. He has contributed to a number of European projects, such as SoBigData and EHRI (European Holocaust Research Infrastructure). His research focuses on digital journalism and its intersection with platforms. Contact: marco.braghieri@kcl.ac.uk

Professor Tobias Blanke is Distinguished University Professor of Humanities and AI at the University of Amsterdam. He has also been Professor in Social and Cultural Informatics and Head of the Department of Digital Humanities at King's College London. His research focuses on big data, AI and their implications for culture and society. Contact: t.blanke@uva.nl

Dr. Jonathan Gray is Lecturer in Critical Infrastructure Studies at the Department of Digital Humanities, King's College London, where he is currently writing a book on data worlds. He is also Cofounder of the Public Data Lab; and Research Associate at the Digital Methods Initiative (University of Amsterdam) and the médialab (Sciences Po, Paris). More about his work can be found at jonathangray.org and he tweets at @jwyg. Contact: jonathan.gray@kcl.ac.uk

References

Albalad Aiguabella, José Maria (2015): Slow journalism para una nueva audiencia digital. El caso de Longform.org (2010-2015). In: *Revista de Comunicación*, 14, pp. 7-25.

Anderson, Chris (2009): *The longer long tail: how endless choice is creating unlimited demand*. New York, NY: Random House Business.

Barker, Deane (2016): *Web Content Management: Systems, Features, and Best Practices*. Sebastopol, CA: O'Reilly Media.

Belt, Don and South, Jeff (2016): Slow Journalism and the Out of Eden Walk. In: *Digital Journalism*, 4(4), pp. 547-562.

Bercovici, Jeff: Longform's New App: More Great Journalism Without The Filter. In: *Forbes*, 17 September 2014. https://www.forbes.com/sites/jeffbercovici/2014/09/17/longforms-new-app-more-great-journalism-without-the-filter/#552df27b42f6 (1 February 2021)

Blanke, Tobias; Prescott, Andrew (2016): Dealing with Big Data. In: Griffin, Gabriele and Hayler, Matt (eds.): *Research Methods for Reading Digital Data in the Digital Humanities*, Edinburgh: Edinburgh University Press, pp. 184-205.

Bødker, Henrik (2018): Journalism History and Digital Archives. In: *Digital Journalism*, 6(9), pp.1113-1120.

Boynton, Robert (2013): Notes toward a Supreme Nonfiction: Teaching Literary Reportage in the Twenty-first Century. In: *Literary Journalism Studies*, 5(2), pp. 125-131.

Brock, George (2013): *Out of print: journalism and the business of news in the digital age*. London: Kogan Page Limited.

Bruns, Axel; Highfield, Tim (2012): Blogs, Twitter, and breaking news: The produsage of citizen journalism. In R.A. Lind (ed.): *Produsing theory in a digital world:*

The intersection of audiences and production in contemporary theory. New York, London: Peter Lang.

Chyi, Hsiang Iris; Chadha, Monica (2012): News on New Devices. In: *Journalism Practice*, 6(4), pp. 431-449.

Coll Ardanuy Mariona; Knauth Jürgen; Beliankou Andrei; van den Bos Maarten; Sporleder Caroline (2016): Person-Centric Mining of Historical Newspaper Collections. In: Fuhr Norbert, Kovács László, Risse Thomas, Nejdl Wolfgang (eds.) *Research and Advanced Technology for Digital Libraries*. TPDL 2016, Lecture Notes in Computer Science, vol. 9819. London: Springer, pp. 320-335.

Data Miner (2016): How Data Miner Works. In: *Data Miner* https://data-miner.io/how-it-works (3 October 2020)

Derrida, Jacques (2005): *Paper machine*. Stanford, CA: Stanford University Press.

Dowling, David; Vogan, Travis (2015): »Can We ›Snowfall‹ This?«. In: *Digital Journalism*, 3(2), pp. 209-224.

Elliott, Chris: The readers' editor on... the richness of digital archives bringing problems for readers and journalists. In: *The Guardian*, 21 October 2012. https://www.theguardian.com/commentisfree/2012/oct/21/digital-archives-problems-readers-journalists (9 January 2021)

Fisher, Mark (2009): *Capitalist realism: is there no alternative?* Alresford: 0 Books.

Franklin, Bob (2014): The Future of Journalism. In: *Journalism Studies*, 15 (5), pp. 481-499.

Hartsock, John C. (2000): *A History of American Literary Journalism: The Emergence of a Modern Narrative Form*. Amherst, MA: University of Massachusetts.

Heyman, Stephen: Google Books: A Complex and Controversial Experiment. In: *The New York Times*, 29 October 2015. https://www.nytimes.com/2015/10/29/arts/international/google-books-a-complex-and-controversial-experiment.html?_r=0 (25 March 2021)

Hiippala, Tuomo (2017): ›The Multimodality of Digital Longform Journalism‹. In: *Digital Journalism*, 5(4), pp. 420-442.

Huang, Sonia J. and Wang, Wei-Ching (2014): Application of the long tail economy to the online news market: Examining predictors of market performance. In: *Journal of Media Economics*, 27(3), pp. 158-176.

Jacobson, Susan; Marino, Jaqueline: Gutsche Robert E Jr; Reynolds, Donald W (2018): Should There Be an App for That? An Analysis of Interactive Applications within Longform News Stories. In: *Journal of Magazine Media*, 18(2).

Jacomy, Mathieu; Venturini, Tommaso; Heymann, Sebastien; Bastian, Mathieu (2014): ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. In: *PLOS ONE*, 9(6), e98679. DOI: https://doi.org/10.1371/journal.pone.0098679

Kachka, Boris: The Future of Reading According to Longform. In: *New York Magazine*, 26 September 2014. http://nymag.com/daily/intelligencer/2014/09/future-of-reading-according-to-longform.html (26 March 2021)
Le Masurier, Megan (2015): What is slow journalism? In: *Journalism practice*, 9(2), pp. 138-152.
Le Masurier, Megan (2016): Slow Journalism: An introduction to a new research paradigm. In: *Journalism Practice*, 10(4), pp. 439-447.
Longform.org (2010): Longform.org. http://www.longform.org (24 June 2020)
Longform.org (2017): Publications. In: Longform.org. https://longform.org/archive/publications?from=1 (24 May 2020)
Longform.org: Why We've Removed Longform from the App Store. In: Longform Medium, 26 September 2014. https://medium.com/@longform/why-weve-removed-longform-from-the-app-store-823d599a34d4 (29 January 2021)
Longhi, Raquel Ritter; Winques, Kérley (2015): The place of longform in online journalism: Quality versus quantity and a few considerations regarding consumption. In: *Brazilian Journalism Research*, 11(1), pp. 104-121.
Marres, Noortje; Weltevrede, Esther (2013): Scraping the Social? In: *Journal of Cultural Economy*, 6(3), pp. 313-335.
Mayer-Schönberger, Viktor; Cukier, Kenneth (2013): *Big data: a revolution that will transform how we live, work and think.* London: John Murray.
McQuade, Eric (2015): Aaron Lammer, The Art of Podcasting. In: *The Timbre*, 20 February 2015. https://web.archive.org/web/20150906075223/http://thetimbre.com/aaron-lammer-art-podcasting-no-5/ (9 December 2020)
Mitchell, Ryan (2015): *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA: O'Reilly Media.
Mullin, B.: New app from Longform allows freelancers to cultivate audiences. In: *Poynter*, 18 September 2014. https://www.poynter.org/news/new-app-longform-allows-freelancers-cultivate-audiences (14 January 2021)
Nel, François; Westlund, Oscar (2012): The 4C's of Mobile News. In: *Journalism Practice*, 6(5-6), pp. 744-753.
Olson, Jack E. (2003): Data Quality: *The Accuracy Dimension*. Burlington, MA: Morgan Kaufmann Publishers.
Pybus, Jennifer; Coté Mark; Blanke, Tobias (2015): Hacking the social life of Big Data. In: *Big Data & Society*, 2(2), pp. 1-10.
Rogers, Richard (2013): *Digital Methods*. Cambridge, MA: MIT Press.
Seaton, Jean (2016): The new Architecture of Communications. In: *Journalism Studies*, 17(7), pp. 808-816.
Shapiro, Michael; Hiatt, Anna; Hoyt, Mike (2015): *Tales From the Great Disruption—Insights and Lessons From Journalism's Technological Transformation*. New York, NY: Big Roundtable Books.

Smith, Virginia; Connor, Miriam; & Stanton, Isabelle (2015): Going In-Depth: Finding Longform on the Web. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2109-2118.

Steensen, Steen; Grøndahl Larsen, Anna M.; Hågvar, Yngve Benestad; Fonn, Brigitte Kjoss (2019): What Does Digital Journalism Studies Look Like? In: *Digital Journalism*, *7*(3), pp. 320–342.

Venturini, Tommaso; Bounegru, Liliana; Gray, Jonathan; Rogers, Richard (2018): A reality check(list) for digital methods. In: *New Media & Society*, 20(11), pp. 4195-4217.

Venturini, Tommaso; Jacomy, Mathieu; Bounegru, Liliana; Gray, Jonathan (2017): Visual network exploration for data journalists. In: Scott Eldridge II and Bob Franklin (eds.): *The Routledge Handbook to Developments in Digital Journalism Studies*. Abingdon: Routledge.

Verborgh, Ruben; De Wilde, Max (2013): *Using OpenRefine*. Birmingham: Packt Publishing.

Yzaguirre, Amelia; Smit, Mike J.; Warren, Rob (2016): Newspaper archives+ text mining= rich sources of historical geo-spatial data. In: *IOP Conference Series: Earth and Environmental Science*, 34(1), pp. 1-8.