A Comparison of News Databases' Coverage of Digital-Native News

Stacy Gilbert and Alexander Watkins

University Libraries, University of Colorado Boulder

Author Note

Correspondence concerning this article should be addressed to Stacy Gilbert, University Libraries: stacy.gilbert@colorado.edu

Abstract

Digital-native news has become widely-read and award-winning sources for news, and it is important to understand if news aggregator databases provide access to these emerging news outlets. This study compares four news aggregators' coverage of popular and Pulitzer Prize finalists digital-native news organizations to print-native news outlets. It found only 14 out of 47 born-digital news organizations are available in the aggregators, and of those outlets, four have a 100% date coverage.

Keywords: aggregators, databases, newspapers, digital news media, online news media, libraries

A Comparison of News Databases' Coverage of Digital-Native News

Introduction

In 2010, ProPublica was the first online news outlet to win a Pulitzer Prize ("2010 Pulitzer Prizes," 2010), but this journalism cannot be found in many news aggregator databases. Born-digital or digital-native news outlets, organizations originally founded on the internet, are some of the most read news websites today. In 1997, 93% of American adults access the news via the web with high-traffic digital-native news outlets seeing 21.7 million monthly unique visitors (Stocking 2018). In comparison, the average monthly unique visitors to the top 50 most circulated U.S. newspapers' websites was 11.5 million (Barthel, 2018).

Not only is born-digital news popular among those seeking news, increasingly, these born-digital news outlets have been investing in quality journalism--recognized by the continued nomination of digital-native outlets for Pulitzer Prizes and other journalism awards. From 2008 to 2018, born-digital news organizations like *Bloomberg News*, *BuzzFeed News*, *HuffPost* (formerly *Huffington Post*), *Reveal*, *InsideClimate News*, *Marshall Project*, and *ProPublica* have either won or were a finalist for a Pulitzer Prize in the news reporting categories ("Prize Winners by Year," 2019). The Online Journalism Awards (2018), Radio Television Digital News Association's National Edward R. Murrow Awards (2018), American Society of Magazine Editors' National Magazine Awards (n.d.), and the Society for Professional Journalists' Sigma Delta Chi Awards (n.d.) have also recognized born-digital news organizations' reporting in the last few years.

Born-digital journalism has been shaping national conversations and breaking major stories, and a significant part of the history of the 21st century is being told in these publications. This study analyzes major news aggregator databases to determine if they are keeping up with

these changes to the media landscape by examining the coverage of born-digital news in news aggregator databases. Since news aggregator databases are frequently used by researchers to discover articles and conduct media analyses, this study has methodological implications for the use of news aggregators in research and content analysis.

Literature Review

News Aggregators and Gatekeeping Theory

News aggregators are databases that aggregate news content for multiple sources for easier searching and browsing. Consumer-facing news aggregators, such as Google News, Yahoo! News, and even Facebook allow users to search for news articles and access the articles by linking to the news article on its publisher's website. If the article is behind a paywall, the consumer will have to pay the publisher to access the article. However, in academia and throughout this article, "news aggregators" refers to paid subscription databases like Access World News, Nexis Uni (formerly LexisNexis Academic), Factiva, and ProQuest U.S. Newsstream. These databases make editorial decisions on which news content to include. They license selected content from newspaper publishers to make available to subscribers in the database until the aggregator-publisher relationship ends. Importantly news aggregators do not serve an archival function, as material frequently comes in and out of the database as licensing agreements change.

Any inconsistent representation of sources in these databases is the result of news aggregators selecting (or not selecting) and licensing publications to include in their databases. This decision-making process can be considered a type of gatekeeping between news articles and scholars. Applicable for understanding knowledge systems, Donohue, Tichenor, and Olien (1972) define gatekeeping as more than just selection of messages but the process involved in

"various forms of information or knowledge control" (p. 42) including "decisions about message encoding, such as selection, shaping, display, timings, withholding, or repetition of entire messages or message components" (p. 43). The gatekeeper is part of the gatekeeping process, making decisions over the information and knowledge control.

Gatekeeping has been typically applied to news professionals' processes for selecting articles for publication (Donohue et al., 1972; Shoemaker & Vos, 2009; White, 1950). With the emergence of various networks created by technology, Barzilai-Nahon (2008) developed network gatekeeping for today's information society. Network gatekeeping theory moves beyond the processes of information selection, distribution, and intermediation by being "a more flexible construct of information control" (Barzilai-Nahon, 2008, p. 1495). It provides the user, or the gated, more power in gated-gatekeeper dynamic by factoring the user has some agency over the information they access. It also accounts for 10 mechanisms that provide a "structure or discourse" to the gatekeeping process, including cost-effect mechanisms (e.g. cost of joining, using, or exiting a network), editorial mechanisms (e.g. technical or content controls), and regulation meta-mechanisms (e.g. agreements, rules, or procedures) (Barzilai-Nahon, 2008, p. 1495).

Helberger, Kleinen-von Königslöw, and van der Noll (2015) used network gatekeeping to examine search engines, social networks, and mobile application stores, arguing that these resources hold characteristics of two types of gatekeeping concepts: "gatekeepers which control access to information and gatekeepers which have a facilitating role through control of critical intermediary resources or services that are necessary to link users and content, mediate between the different players in the information chain" and so on (p. 52). Subscription news aggregator databases also hold these characteristics; they control researchers' access to news articles and

they are intermediary resources that connect users to news content, all while balancing the costs of hosting news outlets' content, the costs of their subscription models, and decisions around which news organizations are important and valuable to users.

Other news aggregation websites and search engines have been studied through various gatekeeper lenses. Algorithmic gatekeeping has been used to discuss personalization features in Google search results and content on Facebook (Bozdag, 2013), and Google News search results during the 2016 U.S. Presidential Election (Nechushtai & Lewis, 2019). Hurley and Tewksbury (2012) compared Google News and Yahoo! News, two popular news aggregation websites at that time, to two news publishers' websites, CNN and MSNBC. They found the "electronic layer of gatekeeping can affect the content and implications of news that audiences receive" (Hurley & Tewksbury, 2012, p. 144).

The Challenges of Research with Born-Digital Sources

Both born-digital and legacy news outlets face challenges archiving and providing access to their content, and they are turning to news aggregators for help. Ringel and Woodall (2019) found legacy and born-digital news outlets "had not given any thought to even basic strategies for preserving their digital content, and not one was properly saving a holistic record of what it produces" (2019, sec. Executive Summary, para. 1). Born-digital news content from small or alternative outlets are at an especially high risk of being lost because these types of organizations can suddenly close and the technologies used to build apps and interactive publishing platforms can become obsolete (Ringel & Woodall, 2019). To help mitigate this problem, some news outlets are using third-party vendors like the Internet Archive, Google, Ancestry.com, and ProQuest to store and provide access to their content (Ringel & Woodall, 2019).

The rapid changes to and disappearance of online sources have made the study of web content particularly challenging. Dimitrova and Bugeja (2007) found that half of the online citations in communications journals no longer worked after two years, calling this phenomenon of rapid decay the half-life of internet references. Web sources have proved especially challenging for content analysis, with McMillan (2000) likening it to using a microscope on a moving target. Challenges identified in the literature include the dynamic nature of web content that is rapidly updated and not published on a pre-planned schedule (McMillan, 200; Karlsson & Sjøvaag 2016; I. Kim & Kuljis 2010). The media-rich, hyper-linked, and interactivtive content of online sources has also posed difficulty for content analysis (I. Kim & Kuljis, 2010; Herring, 2010). Collecting a representative sample can be especially difficult for online sources, especially as such studies may require active and pre-planned data collection to capture digital sources (McMillian, 2000; I. Kim & Kuljis, 2010; Herring, 2010; Karlsson & Sjøvaag, 2016). For many studies news aggregator databases have lessened these problems for online news, especially by archiving stories, allowing studies to analyze web content without pre-planning sample collection.

Content and Media Analyses with News Aggregators

Literature from social sciences, political science, and journalism found that scholars frequently rely on news aggregator databases for conducting studies. Indeed, Deacon (2007) identified eight studies that make use of searching databases to perform content analyses, a number that has likely increased since that article's publication. News articles are used to research a period in time, family history and genealogy, and business dealings (Hansen & Paul, 2015). When systematic data is limited, scholars could use news articles to perform "event count studies" to develop a data set that shows how frequently events like strikes or violence have

occurred (Woolley, 2000). News articles can also be studied to understand the media landscape and its effects on society, such as examining "how journalists construct news articles," or performing a "studies of media focus" which involves measuring how much attention an issue receives by the media over a period of time (Ridout, Fowler, & Searles, 2012, p. 452). For the latter, scholars might look at the number of stories or the number of column inches an issue receives (Atkinson, Lovett, & Baumgartner, 2014; Woolley, 2000).

As the use of news aggregator databases has become increasingly common in media analyses, several studies have raised methodological concerns. One concern is that databases with the same sources often retrieve different articles and a different total number of articles (Blatchford, 2019; Deacon 2007; Sabelhaus & Cawley, 2013; Woolley, 2000). This difference can arise from a number of factors including how the database gathers the content: whether directly from the provider or from web crawling (Blatchford, 2019). Driedger and Weimer (2015) found search results can be inconsistent when searching the same database at different institutions. They discovered Factiva had user preference settings and options that contained exclusion filters. Other items that tend to be missing or have inconsistent coverage in aggregators includes wire service and freelancers' articles, photos and photo captions, illustrations, editorial letters, display and classified advertisements, stock quotations, announcements, meeting calendars, legal notices, and graphs and charts (Orenstein, 1993; Ridout et al., 2012; Snider & Janda, 1998; Weaver & Bimber, 2008). These databases' date coverage can be incomplete. It is not uncommon to find entire years of a newspaper missing from a database. Coverage usually increases as dates get closer to the present, which can create a recency bias (Deacon, 2007). Orenstein (1993) suggests this problem might be due to changing license policies, newspapers' ownership, contract negotiations, name changes, data entry, and so on.

These deficiencies pose challenges to the reliability and verifiability of content analysis studies performed on data gathered from news aggregators. Content analysis is a method for making replicable inferences from texts (Krippendorff, 2013), and the unreliable nature of digital sources has posed problems for the replicability of content analysis performed on online sources (Dimitrova and Bugeja, 2007). Online news databases improved the long-term persistence of these types of online sources, but inconsistencies in search results and content collection still threaten the replicability of studies that rely on these databases (Deacon, 2007). The sampling of sources in a content analysis, generally should be representative and random, with no specific type of source overrepresented (Herring, 2010). The representativeness and randomness of samples drawn from news aggregators is affected by coverage that over or under represents particular dates or types of news articles (Blatchford, 2019). This study looks at the reliability of news databases in representation of born-digital news among the sources they archive and any biases in the date coverage of these sources. Problems of representation of born-digital news poses methodological challenges for studies that rely on these databases as content sources.

Overall, the literature raises questions in news aggregators' gatekeeping role in providing researchers reliable and representative access to born-digital news articles for their scholarship. While news aggregator databases make it easy to search for thousands of articles from online news organizations, and legacy newspapers and magazines, these resources have incomplete coverage of legacy publishing, and the content varies among databases. Due to the popularity of born-digital news, the need for researchers to access this information, and reliance on third-party resources to provide access, a systematic study of born-digital news coverage in news aggregators is needed. The researchers analyzed this topic by asking three research questions:

Q1: What is the coverage of digital-native news outlets available in four popular news aggregator databases: Access World News Research Collection, Factiva, ProQuest U.S. Newsstream, and Nexis Uni?

Q2: How does the coverage of born-digital news outlets compare to print-native news outlets?

Q3: How does the coverage of Pulitzer Prize winning born-digital news outlets compare to Pulitzer Prize winning print-native news outlets?

Methodology

In order to compare digital and print news outlets, the authors found lists of born-digital and award-winning news organizations. For examining major born-digital publications, the authors used Media Cloud's Source Manager's "U.S. Top Online News 2017" (2017) list, which is compiled using data from three website traffic and analytics companies, comScore, Activate, and Alexa, and includes a mix of born-digital news sites and born-print newspapers and magazines. This study also used the "U.S. Top Digital Native News 2016" (2016) list created by the Pew Research Center and comScore, who developed this list based on the number of unique visitors who visited the news outlets' webpages. The authors used these lists to determine the top digital-native news organizations and top print-native news outlets with an online presence. After combining the two lists together, removing duplicate organizations, and omitting wire services and born-broadcast organizations, there were 55 outlets, of which 43 were digital-native and 12 were print-native.

A limitation is while traditional legacy newspapers measure popularity based on subscriptions, born-digital outlets and legacy newspapers' websites measure with web traffic. News organizations employ web analytics to inform how they write headlines, design their

website, and promote stories on social media to attract traffic (Tandoc, 2014; Vu, 2014). As 36% of Americans access news through social media and 20% through search engines (Mitchell et al., 2017), newsrooms that take advantage of search engine optimization, social media algorithms, and clickbait strategies, could affect their standing on these lists. However, the authors determined the "U.S. Top Online News 2017" and "U.S. Top Digital Native News 2016" to be the best lists to use because they were the most current and they described who collected the data.

However, for news organizations high web traffic does not necessarily correlate with quality journalism. To make up for this, the researchers compiled a second list of news and magazine organizations that have won or were finalists for Pulitzer Prizes from 2008 to 2018 in the categories related to news reporting, excluding the editorial, commentary, and photography categories. They also omitted wire services and investigative journalism organizations. There were 77 outlets, 70 of which were print-native and seven were digital-native. After cross-checking the Pulitzer Prize finalists list and the top online and digital-native news lists, the authors removed duplicates and compiled a final list of 121 news outlets for analysis. Forty-seven of these outlets are born-digital and 74 are print-native. A list of the news organizations can be found in Appendix A.

This study analyzes four news aggregators: Access World News Research Collection, Factiva, Nexis Uni, and ProQuest U.S. Newsstream. These databases were selected for a variety of reasons. First, a number of scholarly papers that have analyzed aggregators study these databases or their previous iterations, like Dow Jones; LexisNexis Academic; and ProQuest's Newspaper, Dialog, and Canadian Newsstand Major Dailies (Driedger & Weimer, 2015; Ridout et al., 2012; Sabelhaus & Cawley, 2013; Snider & Janda, 1998; Weaver & Bimber, 2008).

Second, giving variety to the sample, these databases are from three different vendors, with ProQuest U.S. Newsstream and Factiva sharing ProQuest as its vendor. Access World News is owned by NewsBank and Nexis Uni is owned by LexisNexis. Third, while there has been little recent research performed on which databases are available at the most academic libraries, a 2016 study on the databases available at 37 business libraries found LexisNexis Academic was available at 35 libraries, Factiva was in 30 libraries, and NewsBank was in 19 libraries, suggesting these aggregators are popular among researchers and available at numerous universities (K. Kim & Wyckoff, 2016).

In April 2019, the researchers downloaded source lists and dates of coverage from each aggregators' database (LexisNexis, 2018; NewsBank, 2019; ProQuest, 2019; ProQuest LLC, 2019). Using title lists directly from the vendors is similar to Blessinger and Olle's (2004) study which analyzed academic databases with journal content by downloading the title lists from the vendors' websites. The authors used the vendor source lists to compile the coverage dates of each news publication. Since it is not uncommon to find gaps in an aggregators' date coverage of a periodical (Orenstein, 1993), one limitation to this study is the authors only noted the beginning and end dates and did not account for any missing days, months, or years. The authors determined dates of first publication for born-digital publications included in at least one aggregator using publicly available information from the publication's website, news articles covering their launch, and Wikipedia. The authors then compared the total length of publication history of the born digital publications with the aggregator coverage as of April 23, 2019.

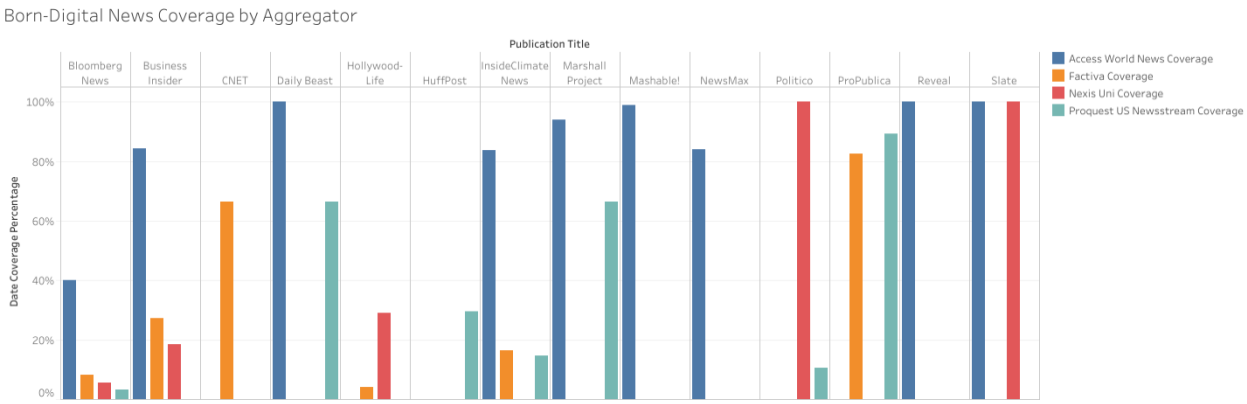
Results

Q1: The Coverage of Digital-Native News Outlets

Of the 47 born-digital news outlets included in the study, only 14 (30%) are available in at least one of the aggregators included in the study. The following titles had at least some coverage in the four aggregators (see Figure 1):

- *Bloomberg News*
- *Business Insider*
- *CNET*
- *The Daily Beast*
- *HollywoodLife*
- *HuffPost* (formerly *The Huffington Post*)
- *InsideClimate News*
- *Marshall Project*
- *Mashable!*
- *NewsMax*
- *Politico*
- *ProPublica*
- *Slate*
- *Reveal* (formerly *California Watch*)

Figure 1



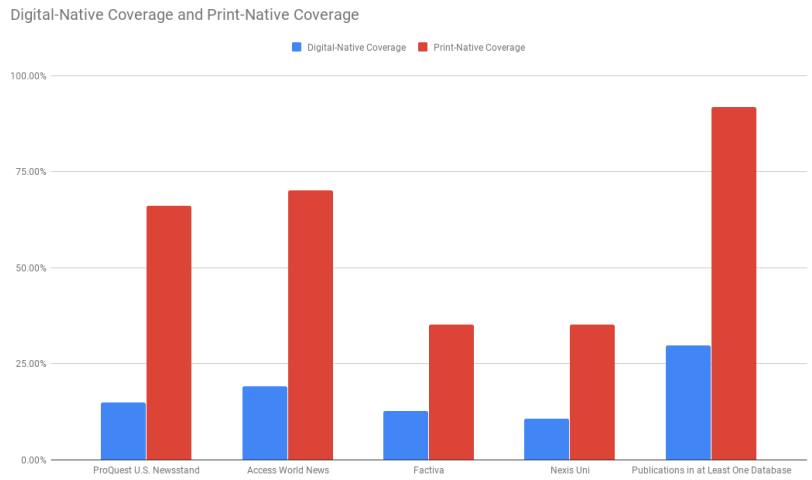
The authors compared the date coverage in aggregators to the publication history of each of these fourteen titles to generate a percentage of content included in each aggregator. Many of the aggregators include only a small percentage of born-digital publications' complete publication dates. Almost all of the publications were included in some but not all aggregators; only *Bloomberg News* was included in all four aggregators. Five publications were included in only one of the aggregators. Access World News has the most date coverage including 100% date coverage of *The Daily Beast*, *Reveal*, and *Slate*, as well as over 80% date coverage of *Business Insider*, *InsideClimate News*, *Marshall Project*, *Mashable!*, and *NewsMax*. Appendix B has a table of the full statistical analysis of the four databases and 14 publications.

Q2: Comparing the Coverage of Born-Digital News Outlets to Print-Native News Outlets

The study found that 30% (n=14) of the born-digital publications included in the study were covered in at least one aggregator. While for born-print publications, 92% (n=68) had coverage in at least one aggregator. Figure 2 compares the percentage coverage of born-digital and print publications in each of the four aggregator databases. Access World News had the highest coverage at 70% of born-print (χ^2 (3) = 32.73, p <.001), and 19% of born-digital publications included in the study (χ^2 (3) = 1.51, n.s). Factiva and Nexis Uni were considerably

lower, both with 35% coverage of born-print titles (χ^2 (3) = 32.73, p <.001). Nexis Uni had the lowest coverage of born-digital titles at 10.6% (χ^2 (3) = 1.51, n.s).

Figure 2

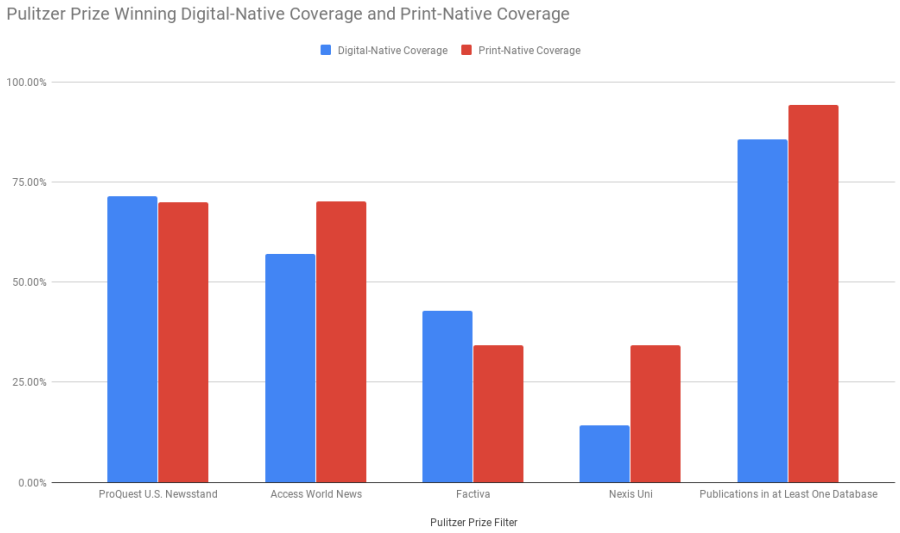


Q3: Coverage of Pulitzer Prize Winning Born-Digital and Print-Native News Outlets

The study found that 94% (n=66) of print-native, Pulitzer Prize finalists had coverage in at least one aggregator, with a significantly higher proportion of the most print-native, Pulitzer finalists' publications in Access World News (70% or n=*51*) and ProQuest (70% or n=49) (χ^2 (3) = 38.87, p <.001). The four print-native Pulitzer Prize finalist publications that cannot be found in one of the databases are *Advertiser Democrat* (Maine), *Concordia (La) Sentinel*, *National Geographic*, and the *Stranger* (Seattle). The percentage was slightly smaller, 86% (n=6), for digital-native, Pulitzer Prize finalists. Of the born-digital publications, only *BuzzFeed News* did not have any coverage in one of the four aggregator databases. Figure 3 compares the percentage coverage of born-digital and print publications that have been Pulitzer Prize finalists in each of the four aggregator databases. ProQuest has the most coverage of digital-native, Pulitzer Prize finalists (71.4% or n=5) (χ^2 (3) = 5.03, n.s.), despite having less overall born-

digital coverage than Access World News. Nexis Uni had the lowest coverage of born-digital, Pulitzer Prize finalist publications at 14.3% (n=1) (χ^2 (3) = 5.03, n.s.).

Figure 3



Discussion

Increasingly, born-digital news organizations are telling the story of the 21st century; however, this study found most born-digital news content is not accessible through the major news databases. Only 30%, or 14 out of 47, born-digital news outlets have content in one of the four news aggregators. Of the 14 born-digital news organizations, only four publications, *The Daily Beast*, *Reveal*, *Slate*, and *Politico*, have 100% date coverage in either Access World News and/or Nexis Uni. News aggregators have become 21st century gatekeepers, by selecting the publications that will be included and excluded from these core databases. They exclude 70% of the most read or award-winning born-digital news outlets, which has implications for researchers who rely on aggregators to assess coverage of events in the 21st century. Research will be missing key articles from born-digital publications that are breaking stories and shaping the media landscape. Sampling undertaken using these databases as a data source have the potential

to underrepresent born-digital sources. This could skew studies of media attention, especially as born-digital publications may have differing content focuses then their print counterparts.

There are also large gaps in date coverage of born-digital publications. News databases are just now beginning to add coverage of fundamental born-digital publications and most are not adding content retrospectively. This lack of comprehensive historic coverage for born-digital publications represents a major lacuna in the historical record of the early 21st century. News organizations who are relying on news aggregators to make their historical content accessible should be aware of the gaps that might persist in working with aggregators and determine short and long-term solutions for ensuring complete coverage. Researchers should be aware that just because more recent issues of born-digital publications may be available in a database, historical coverage is often missing, and this could affect their study results by creating a recency bias.

Conversely, coverage of born-digital, Pulitzer Prize finalists are high in aggregators. The authors can speculate that the coverage of born-digital, non-Pulitzer Prize finalist organizations is low due to a perception of low-quality journalism among born-digital publications, an instance of gatekeeping resulting from news aggregators making selection decisions. An illustrative example is *BuzzFeed News,* which is not included in any aggregator and whose parent website, *BuzzFeed*, is known most for listicles and personality quizzes. In 2014, few Americans were aware of its news division or distrusted their reporting (Pew Research Center, 2014, p. 17). However, *BuzzFeed News* has broken important stories and has been a Pulitzer finalist in 2016 and 2017 ("20 Facts You Probably Didn't Know About BuzzFeed News," 2018). As born-digital sources continue to create high-quality, award-winning journalism and play leading roles in shaping the media and news landscape, perhaps coverage for all born-digital sources will improve.

In addition to *BuzzFeed News*, four print-native, Pulitzer Prize finalists, *Advertiser Democrat* (Maine), *Concordia (La) Sentinel*, *National Geographic*, and *Stranger* (Seattle) are not available in any database. Researchers should be aware that a search in a news aggregator might be missing important national and global news from *BuzzFeed News* and *National Geographic*. As local news offices merge and shrink, researchers may not be able to access key stories with regional roots, as illustrated by the lack of coverage of the *Advertiser Democrat*, *Concordia (La) Sentinel*, and *Stranger*.

Best Practices and Recommendations

Through this research, the authors have identified a number of best practices scholars should consider when conducting a news analysis using news aggregators. As each news aggregator carries different publications and date coverage, scholars should research which publications and date ranges they would like to include in their sample and then choose an aggregator that fits their methodology. If projects are tied to using only one specific database, they should review the aggregator's content lists for gaps to understand any limitations to their methodology. Researchers should consider using multiple aggregators for a more complete coverage. For example, although Access World News Research Collection had the most born-digital news publications, an informal review of the database shows it lacks major news organizations like *The New York Times*, *The Los Angeles Times*, and *The Washington Post*, which are available in Nexis Uni. Studies that use more than one aggregator may increase the comprehensiveness of born-digital and born-print content found. Additionally, researchers should be aware that currently even using multiple databases will not search born-digital content in a comprehensive manner. They may need to resort to other methods such as web scraping or freely available search engines like Google News or Bing News to supplement their studies.

However, these different tools will often provide inconsistent number results and archived articles, making it challenging to achieve consistent, replicable results that can be compared across databases (Blatchford, 2019).

There are a number of ways news aggregator databases and news publications can improve their content offerings to better serve users, however, the authors recognize the gatekeeping processes of news aggregators providing access depends on various mechanisms, namely costs of adding publications and copyright laws (Barzilai-Nahon, 2008). To start, news aggregators should stay abreast with the latest news content producers by monitoring the news outlets Americans frequently read, and offer this content in their products. If the licensing costs allow, news aggregators should add born-digital publications that are significantly impacting public discourse and the news landscape. This may be especially important for Nexis Uni as LexisNexis (2017, para. 4) lauded Nexis Uni's ability to "quickly sift through countless websites and databases and return a clear visual presentation of relevant results." Despite this, Nexis Uni had the smallest amount of digital-native news content and was tied with Factiva for the fewest print-native coverage sources. As news organizations and the ways people access news evolve, so too should news aggregators' content.

Lastly, as scholars often construct systematic studies around aggregator searches, it would be useful if aggregators provided transparent information around the type of content included and excluded from their databases. In addition to a content list, databases could provide a summary of its strengths and weaknesses regarding the types of news articles and sources they offer. Information around date ranges and other gaps, such as exclusion of wire services, freelancers' articles, photos, and other news features, would also be useful. Understanding the strengths and limitations of aggregators can help researchers select databases for their studies.

Conclusion

News aggregator databases have the potential to assist scholars in their web content research, however, the authors found that very few of the United States' most read born-digital news organizations are available in news aggregators. As these news aggregators are one of the few ways to analyze online news without years-long data collection, the lack of born-digital news represents a kind of network gatekeeping. This gatekeeping has reliability and verifiability implications for scholars who depend on news aggregator databases for content analysis research as it will skew sampling and affect replicability. Researchers will miss high-quality and frequently read articles that have a social impact. Furthermore, this is a cause of concern for news outlets that rely on third-party aggregators to make their content accessible for future use. This study raises several questions for future research. While this study analyzed missing date coverage, it would also be advantageous to see if aggregators have other gaps in born-digital outlets' content, such as photos, charts and graphs, videos, and interactive features, particularly since these gaps are found in print sources. It would also be worthwhile to research if entire articles are included or excluded from databases, or which editions of born-digital articles (for example, different iterations of breaking news pieces) are accessible in aggregators. Increasingly these kinds of features are equally as important as the text for the news story, especially as data journalism continues to increase in importance. This study was also limited to born-digital and print organizations, leaving broadcast and wire services unstudied. Future research should analyze the coverage of these organizations' content.

Bibliography

20 Facts You Probably Didn't Know About BuzzFeed News. (2018, November 19). *BuzzFeed News*. Retrieved from https://www.buzzfeednews.com/article/buzzfeednews/buzzfeed-news-facts

2010 Pulitzer Prizes. (2010, April 12). Retrieved July 12, 2019, from The Pulitzer Prizes website: https://www.pulitzer.org/prize-winners-by-year/2010

2018 Online Journalism Awards Finalists. (2018). Retrieved April 29, 2019, from Online Journalism Awards website: https://awards.journalists.org/winners/2018/

Atkinson, M. L., Lovett, J., & Baumgartner, F. R. (2014). Measuring the media agenda. *Political Communication*, *31*(2), 355–380. https://doi.org/10.1080/10584609.2013.828139

Barthel, M. (2018, June 13). Newspapers Fact Sheet. Retrieved April 29, 2019, from Pew Research Center website: https://www.journalism.org/fact-sheet/newspapers/

Barzilai-Nahon, K. (2008). Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American Society for Information Science and Technology*, *59*(9), 1493–1512. https://doi.org/10.1002/asi.20857

Blatchford, A. (2019). Searching for online news content: The challenges and decisions. *Communication Research and Practice*, 1–14. https://doi.org/10.1080/22041451.2019.1676864

Blessinger, K., & Olle, M. (2004). Content analysis of the leading general academic databases. *Library Collections, Acquisitions, & Technical Services*, *28*(3), 335–346. https://doi.org/10.1080/14649055.2004.10766000

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology, 15*(3), 209-227. https://doi.org/10.1007/s10676-013-9321-6

Deacon, D. (2007). Yesterday's papers and today's technology: Digital newspaper archives and 'push button' content analysis. *European Journal of Communication*, *22*(1), 5–25. https://doi.org/10.1177/0267323107073743

Dimitrova, D. V., & Bugeja, M. (2007). The half-life of internet references cited in communication journals. *New Media & Society*, *9*(5), 811–826. https://doi.org/10.1177/1461444807081226

Donohue, G. A., Tichenor, P. J., & Olien, C. L. (1972). Gatekeeping: Mass media systems and information control. In F. G. Kline & P. J. Tichenor (Eds.), *Current Perspectives in Mass Communication Research* (pp. 165–174). Beverly Hills, CA: SAGE.

Driedger, S. M., & Weimer, J. (2015). Factiva and Canadian Newsstand Major Dailies: Comparing retrieval reliability between academic institutions. *Online Information Review*, *39*(3), 346–359. https://doi.org/10.1108/OIR-11-2014-0276

Hansen, K. A., & Paul, N. (2015). Newspaper archives reveal major gaps in digital age. *Newspaper Research Journal*, *36*(3), 290–298. https://doi.org/10.1177/0739532915600745

Helberger, N., Kleinen-von Königslöw, K., & van der Noll, R. (2015). Regulating the new information intermediaries as gatekeepers of information diversity. *Info*, *17*(6), 50–71. https://doi.org/10.1108/info-05-2015-0034

Herring, S. C. (2010). Web content Analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 233–249). https://doi.org/10.1007/978-1-4020-9789-8_14

Hurley, R. J., & Tewksbury, D. (2012). News aggregation and content differences in online cancer news. *Journal of Broadcasting & Electronic Media*, *56*(1), 132–149. https://doi.org/10.1080/08838151.2011.648681

Karlsson, M., & Sjøvaag, H. (2016). Content analysis and online news. *Digital Journalism*, *4*(1), 177–192. https://doi.org/10.1080/21670811.2015.1096619

Kim, I., & Kuljis, J. (2010). Applying content analysis to web-based content. *Journal of Computing and Information Technology*, *18*(4), 369–375. https://doi.org/10.2498/ cit.1001924

Kim, K., & Wyckoff, T. (2016). What's in your list?: A survey of business database holdings and funding sources at top academic institutions. *Journal of Business & Finance Librarianship*, *21*(2), 135–151. https://doi.org/10.1080/08963568.2016.1140548

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed). Los Angeles; London: SAGE.

LexisNexis. (2017, January 19). LexisNexis Launches New Academic Research Solution Designed for Millennials. Retrieved May 20, 2019, from LexisNexis website: https://www.lexisnexis.com/en-us/about-us/media/press-release.page?id=1484689760846412&y=2017

LexisNexis. (2018, September). Nexis Uni Support & Training. Retrieved April 9, 2019, from LexisNexis website: https://www.lexisnexis.com/en-us/support/nexis-uni/default.page

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism and Mass Communication Quarterly, 77*(1), 80–98. https://doi.org/10.1177/ 107769900007700107

Mitchell, A., Gottfried, J., Shearer, E., & Lu, K. (2017, February 9). *How Americans encounter, recall and act upon digital news*. Pew Research Center's Journalism Project. https://www.journalism.org/2017/02/09/how-americans-encounter-recall-and-act-upon-digital-news/

National Magazine Award Winners 1966-2015. (n.d.). Retrieved April 29, 2019, from American Society of Magazine Editors website: https://asme.magazine.org/asme/national-magazine-award-winners-1966-2015#Reporting

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298–307. https://doi.org/10.1016/j.chb.2018.07.043

NewsBank. (2019). Browse Publication. Retrieved April 10, 2019, from Access World News website: https://infoweb-newsbank-com.colorado.idm.oclc.org/apps/news/source-list?p=AWNB

Orenstein, R. M. (1993). "How full is full" revisited: A status report on search full-text periodicals. *Database*, *16*(5), 14–23.

Pew Research Center. (2014). *Political Polarization & Media Habits: From Fox News to Facebook, How Liberals and Conservatives Keep Up with Politics* (p. 81). Retrieved from https://www.pewresearch.org/wp-content/uploads/sites/8/2014/10/Political-Polarization-and-Media-Habits-FINAL-REPORT-7-27-15.pdf

Prize Winners by Year. (2019). Retrieved April 29, 2019, from The Pulitzer Prizes website: https://www.pulitzer.org/prize-winners-by-year

ProQuest. (2019, March 20). Factiva: About. Retrieved April 9, 2019, from ProQuest website: http://proquest.libguides.com/factiva/about

ProQuest LLC. (2019, March 30). Title Lists System. Retrieved April 9, 2019, from ProQuest website: http://tls.search.proquest.com/titlelist/jsp/list/tlsSingle.jsp?productId=10000267

Ridout, T. N., Fowler, E. F., & Searles, K. (2012). Exploring the validity of electronic newspaper databases. *International Journal of Social Research Methodology*, *15*(6), 451–466. https://doi.org/10.1080/13645579.2011.638221

Ringel, S., & Woodall, A. (2019). *A Public Record at Risk: The Dire State of News Archiving in the Digital Age*. Retrieved from The Tow Center for Digital Journalism at Columbia's Graduate School of Journalism website: https://www.cjr.org/tow_center_reports/the-dire-state-of-news-archiving-in-the-digital-age.php

RTDNA Announces 2018 National Edward R. Murrow Awards. (2018, June 19). Retrieved April 29, 2019, from Radio Television Digital News Association website: https://rtdna.org/article/rtdna_announces_2018_national_edward_r_murrow_award_winners

Sabelhaus, L., & Cawley, M. (2013). Searching for news online: Challenging traditional methods. *Online Searcher*, *37*(2), 10–14.

Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. New York: Routledge.

Sigma Delta Chi Awards. (n.d.). Retrieved April 29, 2019, from Society of Professional Journalists website: https://www.spj.org/sdxa18.asp

Snider, J. H., & Janda, K. (1998). *Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis*. 29. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.502.7296&rep=rep1&type=pdf

Stocking, G. (2018, June 6). Digital News Fact Sheet. Retrieved April 26, 2019, from Pew Research Center website: https://www.journalism.org/fact-sheet/digital-news/

Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping: *New Media & Society*. https://doi.org/10.1177/1461444814530541

U.S. Top Digital Native News 2016. (2016). Retrieved March 19, 2019, from Media Cloud website: https://sources.mediacloud.org/#/collections/57078150

U.S. Top Online News 2017. (2017, August). Retrieved March 19, 2019, from Media Cloud website: https://sources.mediacloud.org/#/collections/58722749

Vu, H. T. (2014). The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism: Theory, Practice & Criticism*, *15*(8), 1094–1110. https://doi.org/10.1177/1464884913504259

Weaver, D. A., & Bimber, B. (2008). Finding news stories: A Comparison of searches using LexisNexis and Google News. *Journalism & Mass Communication Quarterly*, *85*(3), 515–530. https://doi.org/10.1177/107769900808500303

White, D. M. (1950). The "Gate Keeper": A case study in the selection of news. *Journalism Quarterly*, *27*(4), 383–390. https://doi.org/10.1177/107769905002700403

Woolley, J. T. (2000). Using media-based data in studies of politics. *American Journal of Political Science*, *44*(1), 156–173. https://doi.org/10.2307/2669301

Appendix

Appendix A: List of Titles Analyzed

- 247Sports
- Atlantic, The
- Bleacher Report
- Bloomberg News
- BuzzFeed
- Huffington Post (HuffPost)
- Bustle
- Breitbart
- CNET
- Daily Beast, The
- Business Insider
- Deadspin
- digitaltrends.com
- Daily Caller, The
- Elite Daily
- FiveThirtyEight
- Forbes
- Gizmodo
- Guardian, the - United States
- Hello Giggles
- HollywoodLife

- Drudge Report
- Iflscience
- IGN
- Independent Journal Review (IJR)
- International Business Times, The (IBTimes.com)
- Los Angeles Times, The
- Mashable!
- Mic
- New York Daily News
- New York Times, The
- New Yorker, The
- NewsMax
- Opposing Views
- Politico
- Quartz
- Raw Story
- Refinery29
- SB Nation
- Slate.com
- TheBlaze
- Thrillist
- Time
- TMZ

- Topix
- Uproxx
- Upworthy
- USA Today
- Verge
- Vox
- Wall Street Journal
- Wall Street Journal Blogs
- Washington Post
- Yahoo Finance
- Yahoo News - Latest News & Headlines

Appendix B: Table 1 Born-Digital News Coverage by Aggregator with Statistical Analysis

Table 1: Born-Digital News Coverage by Aggregator with Statistical Analysis						
	Number of Days Articles from Publications are Found in Each Aggregator				Total Number of Days Publisher has Existed	Proportions Test
Publications	**ProQuest US Newsstream**	**Access World News**	**Factiva**	**NexisUni**		
Bloomberg News	280	3550	729	489	8878	$\chi 2$ (3) = 6540.6, <.001
Business Insider	0	3672	1191	799	4360	$\chi 2$ (3) = 7872.7, <.001
CNET	0	0	5652	0	8513	$\chi 2$ (3) = 20330, <.001
Daily Beast	2563	3851*	0	0	3851	$\chi 2$ (3) = 11876, <.001
Hollywood-Life	0	0	158	1090	3764	$\chi 2$ (3) = 2878.6, <.001
HuffPost	1510	0	0	0	5097	$\chi 2$ (3) = 4892.3, <.001
InsideClimate News	657	3764	736	0	4495	$\chi 2$ (3) = 9236.2, <.001
Marshall Project	1147	1620	0	0	1726	$\chi 2$ (3) = 4887.6, <.001
Mashable!	0	4978	0	0	5026	$\chi 2$ (3) = 19849, <.001
NewsMax	0	6321	0	0	7524	$\chi 2$ (3) = 24005, <.001
Reveal	0	3511	0	0	3511	$\chi 2$ (3) = 14044, <.001
Politico	477	0	0	4473**	4473	$\chi 2$ (3) = 15763, <.001
ProPublica	4016	0	3703	0	4495	$\chi 2$ (3) = 13570, <.001
Slate	0	8337	0	8337	8337	$\chi 2$ (3) = 33348, <.001
Proportions Test	$\chi 2$ (3) = 41663, <.001	$\chi 2$ (3) = 55940, <.001	$\chi 2$ (3) = 38729, <.001	$\chi 2$ (3) = 62496, <.001		
	* Access World News has 3852 days of the Daily Beast's coverage, or >100%, which the authors reduced to 3851 to find $\chi 2$ ** Nexis Uni has 4475 days of *Politico*'s coverage, or >100%, which the authors reduced to 4473 to find $\chi 2$					