Volume 6 Number 7 September 2018

DIGITAL JOURNALISM

Routledge

Digital Journalism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rdij20

The Invisible Infrastructures of Online Visibility: An Analysis of the Platform-Facing Markup Used by U.S.-Based Digital News Organizations

Bernat Ivancsics, Eve Washington, Helen Yang, Emily Sidnam-Mauch, Ayana Monroe, Errol Francis II, Joseph Bonneau, Kelly Caine & Susan E. McGregor

To cite this article: Bernat Ivancsics, Eve Washington, Helen Yang, Emily Sidnam-Mauch, Ayana Monroe, Errol Francis II, Joseph Bonneau, Kelly Caine & Susan E. McGregor (2023): The Invisible Infrastructures of Online Visibility: An Analysis of the Platform-Facing Markup Used by U.S.-Based Digital News Organizations, Digital Journalism, DOI: 10.1080/21670811.2022.2156365

To link to this article: https://doi.org/10.1080/21670811.2022.2156365

DIGITAL JOURNALISM
https://doi.org/10.1080/21670811.2022.2156365

The Invisible Infrastructures of Online Visibility: An Analysis of the Platform-Facing Markup Used by U.S.-Based Digital News Organizations

Bernat Ivancsics[a], Eve Washington[a], Helen Yang[a], Emily Sidnam-Mauch[b], Ayana Monroe[c], Errol Francis II[b], Joseph Bonneau[d], Kelly Caine[b] and Susan E. McGregor[a]

[a]Columbia University, New York, NY, USA; [b]Clemson University, Clemson, SC, USA; [c]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [d]New York University, New York, NY, USA

ABSTRACT

This study analyzes and compares how the digital semantic infrastructure of U.S. based digital news varies according to certain characteristics of the media outlet, including the community it serves, the content management system (CMS) it uses, and its institutional affiliation (or lack thereof). Through a multi-stage analysis of the actual markup found on news outlets' online text articles, we reveal how multiple factors may be limiting the discoverability and reach of online media organizations focused on serving specific communities. Conceptually, we identify markup and metadata as aspects of the semantic infrastructure underpinning platforms' mechanisms of distributing online news. Given the significant role that these platforms play in shaping the broader visibility of news content, we further contend that this markup therefore constitutes a kind of infrastructure of visibility by which news sources and voices are rendered accessible—or, conversely—invisible in the wider platform economy of journalism. We accomplish our analysis by first identifying key forms of digital markup whose structured data is designed to make online news articles more readily discoverable by search engines and social media platforms. We then analyze 2,226 digital news stories gathered from the main pages of 742 national, local, Black, and other identity-based news organizations in mid-2021, and analyze each for the presence of specific tags reflecting the Schema.org, OpenGraph, and Twitter metadata structures. We then evaluate the relationship between audience focus and the robustness of this digital semantic infrastructure. While we find only a weak relationship between the markup and the community served, additional analysis revealed a much stronger association between these metadata tags and content management system (CMS), in which 80% of the attributes appearing on an article were the same for a given CMS, regardless of publisher, market, or audience focus. Based on this finding, we identify the organizational characteristics that may influence the specific CMS used for digital publishing, and, therefore, the robustness of the digital semantic

KEYWORDS

Digital journalism; metadata; infrastructure; Schema.org; platforms; local news; ethnic news

CONTACT Bernat Ivancsics bi2155@columbia.edu; Susan E. McGregor sem2196@columbia.edu

infrastructure deployed by the organization. Finally, we reflect on the potential implications of the highly disparate tag use we observe, particularly with respect to the broader visibility of online news designed to serve particular US communities.

Introduction

Although it arguably represents the most important part of an online news article, the visible content presented to audiences on a news outlet's website is just one part of a complex ecosystem of digital infrastructure that influences not *only what* audience members see when they read a news article online, but *whether* they ever encounter that content in the first place. Beneath the visible text and media content on news websites, a wide range of *metadata* is embedded in the markup of the news web pages, some of it designed specifically to make the content legible to the search engines and social media platforms upon which so many news publishers depend to generate visibility for their journalism and, in many cases, revenue.

In this study, we collect a broad sampling of articles from U.S. news organizations that publish on the Web and extract select markup tags from their text (as opposed to audio- or video-focused) article web pages. We then analyze the variety and prevalence of these tags and their attributes through the conceptual framework of the "digital semantic infrastructure." By treating this markup layer as an essential structural aspect of news organizations' online representation and discoverability, we identify and elaborate the relationships between news organizations' audience focus and resources and the robustness of their search-engine and platform-facing digital markup. In conceptualizing markup as part of news organizations' digital infrastructure, this work foregrounds the networked nature of digital news, and situates digital markup practices within the broader ecosystem of "platform journalism" (Burgess and Hurcombe 2019; Thorson et al. 2020). By revealing patterns and discrepancies in markup deployment across publications with different target audiences and resource profiles, this work reveals the markup employed by news organizations and how these markup practices, in turn, affect publishers' visibility to audiences given the platform-driven, algorithmically curated information economy (Van Dijck, Poell, and De Waal 2018). In short, our inquiry seeks to assess the degree to which digital news publishers support and populate different types of platform-facing markup, as well as identify the degree to which patterns and disparities in markup usage correlate with news organizations' other characteristics, such as target audience, technology stack and network membership—or lack thereof. In examining these relationships, this work reveals that far from "leveling the playing field" of news publishing and distribution, the rise of web-based digital publishing appears to be replicating print-era disparities in news organizations' reach, with independent and minority-focused publications under-utilizing the structured data required to make their web-based publications visible to distribution platforms and, consequently, internet audiences in general.

To begin answering these questions, we first review common types of structured data used in US and global news production, and select specific tags designed to

support indexing by search engines and social media platforms as the focus of our analysis. We then map and compare the markup practices of U.S.-based news organizations (Mika 2015; Sunne 2021) according to their audiences of focus—specifically national, geographically regional, Black, and ethnic and/or diaspora communities. While we discover a weak relationship between audience and markup robustness, additional analysis revealed a much stronger relationship between news organizations' content management system (CMS) and the variety and completeness of discoverability-related markup present on article web pages. Pursuing this further, we find that massive discrepancies in the cost and availability of CMSs themselves may help explain why news organizations serving audiences underrepresented in the journalism industry may continue to be less visible than their larger, better-resourced and more interconnected counterparts, despite the free and open-source nature of digital markup standards. As a result, these publications are likely at a disadvantage in terms of gaining recognition and revenue for their work, due to their reduced visibility to search engine and social media platforms' automatic curation systems.

Background and Literature Review

As search engines and social media platforms have replaced professional news organizations as the de facto curators of public information over the past twenty years, new digital practices and formats have emerged as news publishers adapt to their changing role in society and public discourse. While some of these changes have been content-related (including the increased reliance on content formats that are by design "shareable" in social media feeds), many of them apply to the digital infrastructure of online publishing and curation whose semantics are expressed and embodied in a range of digital markup, which is designed to be legible principally to web technologies and platforms, and which typically go unseen by human audiences. Yet because of the role of this digital markup layer as an interface between web technologies—particularly search engines and social media platforms—and news publishers, we argue that investigating the prevalence and robustness of certain markup structures across news web pages can provide valuable insight into the functioning of the online news publishing ecosystem.

A New Focus on Markup as Digital Semantic Infrastructure

This study situates the platform-facing mechanisms of distributing digital news—including the transmission protocols, file formats, and indexing mechanisms that govern how digital content is circulated across the online ecosystem (Fensel et al. 2001)—as part of the code- and convention-based "semantic infrastructure" that plays a pivotal role in the material visibility of digitally distributed news content. As such, we contend that this collection of technologies also constitute an "infrastructure of visibility," which manifests in the interaction between news organizations' structured markup and platform companies' curation algorithms. Given the central role of those platforms in surfacing content to users, robust markup becomes an essential

component of the visibility of online news—even as the metadata layers that drive this process are largely invisible to most human users.

Key parts of the semantic infrastructure of the Web are implemented as metadata-bearing *markup*, where structured "meta" information is included (but often not visibly rendered) alongside the core, visible "content" of a web page. We note here that the term "infrastructure" is typically used to invoke mediating technologies that comprise "neutral" surfaces and foundations for basic utility functions in society (Bowker and Star 1998), but that while this characterization can help build consensus around the appropriateness of equitable access to those technologies, they can also deflect attention from the circumstances that lead to disparate access to—and value derived from—those technologies in practice. In online news publishing, for example, the "free and open source" (FOSS) nature of the markup standards comprising the digital semantic infrastructure used by search engines and social media platforms may mask differences in news publishers' ability to implement this markup and therefore benefit fully from the semantic infrastructure that markup embodies. These markup structures—whose definitions originate from a variety of organizations and interest groups, media organizations and technology companies among them—reflect the degree to which digital news providers are participants in a networked system of institutions and knowledge domains (Ananny 2018; Christin 2020).

Previous literature thematized the infrastructural aspects of this semantic layer only tangentially. Scholarship investigating the development of the semantic web (Schwartz 2003; Hitzler 2021) and the materiality of the public internet's infrastructure throughout the 1990s and early 2000s (Dowd 2020) has primarily focused on the implementations of storage and transmission technologies, such as mainframes, cables, and other hardware. Additionally, scholars have investigated how news organizations craft their content specifically for algorithmic news distribution (Bandy and Diakopoulos 2020), and they have also explored the ways in which news organizations have worked to optimize their distribution mechanisms for search engines, social media platforms, or mobile devices, albeit with no particular focus on metadata or markup (Ananny and Crawford 2015). When scholars did propose specific markup changes with the goal of better catering to digital publishers' needs, they took the platforms' point of view (Kodama, et al. 2008).

Additionally—and to further situate the current analysis at the juncture of platform and journalism studies—, scholarship problematizing the convergence of news publishing with distribution platforms has addressed a range of further issues, including social media platforms' impact on the relationship between news media and their audiences (Bucher 2017; Erwig 2017); the compatibility of mobile devices and news applications (Ananny and Crawford 2015; Kammer 2021); the effects of news feed and news app personalization (Schjøtt Hansen and Hartley 2021); and audience reciprocity in the production of digital news content (Belair-Gagnon, Nelson, and Lewis 2019). Considerable research has also focused on search engines' and social media platforms' algorithmic approach to curation (Diakopoulos 2015) as well as news websites' efforts to optimize their content for these algorithms (Diakopoulos 2020).

Yet while some recent research has explored the relationship between markup quality and information quality (Kennedy and Griffith 2020; Castelo, et al. 2019), there has

been almost no work examining the semantic infrastructure of digital news pages that serves as an essential part of the input to these algorithms. Especially given recent scholarship indicating the diminished representation within the news content surfaced algorithmically by news feeds and content aggregators (Sherry 2015; Sherry and Matsaganis 2019), more attention to journalistic markup practices and their implications is warranted. The current work seeks to help fill that gap by providing an empirical assessment of contemporary markup practices and evaluating its potential relationship to the discoverability and reach of digital news targeting specific audiences in the US.

Markup in Digital News Publishing

Although early digital markup structures for news evolved alongside the Web, their initial purpose was to support the digital transmission of multimedia artifacts and accompanying metadata within and among news agencies and wire services, and therefore also support the "creation, editing, management, and publication of news in a networked computing environment" (*IPTC NewsML 1.0 Specification* 2002). While the NewsML1 protocol introduced in 2000 by the International Press and Telecommunications Council (IPTC) therefore focused on the needs of news publishers, when NewsML2 made its debut in 2008, its specification was designed through a collaboration among news agencies, web browsers (including Safari and Firefox), and system vendors and news aggregators—including Google News and Flipboard. In addition to being interoperable with web-standard HTML and XML, NewsML2 illustrates both the specificity and complexity of digital semantic infrastructures: mandatory attributes include discrete document identifiers, version numbers, publishing status, embargo information, correction signals, and intellectual property identifiers. Although conceptually legible to search engines and social media platforms, however, NewsML2's greatest application may be in Agence France Presse's NewsML-G2 document format, which is primarily used to share multimedia content among international wire services (*IPTC NewsML G2.0 Specification* 2021).

In 2010, Facebook introduced the OpenGraph protocol (Recordon 2010), the design of which was informed by publisher concerns around the impact on search ranking of including certain types of information in existing HTML metadata tags. At the same time, a core focus of OpenGraph is to support an abbreviated rendering of news content in the Facebook news feed (*OpenGraph Markup Guide* 2021), a goal shared by the specific structure of Twitter markup tags and attributes, which supports the rendering of article "cards" in a Twitter timeline (*Twitter Cards Guide* 2021).

In June 2011, search engines Google, Microsoft, Yahoo and Yandex launched the Schema markup structure via Schema.org (Mika 2015; Guha, Brickley, and Macbeth 2016). Though not applicable exclusively to news content, the news-relevant structures of the Schema specification draw heavily on NewsML. At the same time, it is designed principally to embed metadata into audience-facing web pages in order to support support indexing by search engines. While maintained collaboratively through the W3C Schema.org Community Group, the steering committee includes representatives from platform companies including Google and Microsoft (*Schema.org History of Development* 2015).

While concerns about efforts to "game" platform companies' algorithms limits the detail they will provide as to the specific role and significance of any particular web page feature—even and including markup conforming to standards that they themselves have defined—it is widely known that the Twitter platform will make use of OpenGraph structured data for indexing in the absence of its own specific tags. Google, meanwhile, indicates on its developers page that it only supports structured data using the Schema.org vocabulary (*General structured data guidelines* 2022).

Given the long-standing nature of these markup specifications and their pivotal role in the distribution of digital news, the goal of this work is to assess the degree to which news publishers serving various audiences in the US are leveraging the digital semantic infrastructure provided by these markup specifications. In order to understand the role that this markup plays in the current ecosystem of digital news publishing, we therefore formulate the following two research questions:

> RQ1: To what degree do US news organizations with an online publishing presence employ platform-facing markup on their text-based news articles?
>
> RQ2: What association, if any, exists between the characteristics of a news organization—such as its size, network membership, and type of audience served—and the degree or robustness of structured data supporting its digital articles?

By understanding differences in adoption and implementation of these metadata structures, we will shed light on the degree to which these "neutral" foundations are resulting in disparate discoverability across digital news organizations in the US.

Methodology

In this section, we review the design of our data collection process and focus on our selection of relevant markup structures and sample of online news publishers, as well as on the pilot and final sample selection of actual news articles for analysis. We also review our ethics considerations regarding the automated collection of content from online news publishers, and how we sought to minimize any possible negative effects of our process.

Markup Selection and Availability

In order to assess the digital semantic infrastructure of online news publishers, our team began by reviewing the markup protocols discussed in section 2.2. Given that the various NewsML protocols we reviewed are largely implemented as business-to-business solutions among wire services, we chose to focus on the platform-facing markup structures described by Schema.org, Facebook OpenGraph, and Twitter "card" metadata. Through a brief review of the markup available via the common "View Page Source" context menu item on leading web browsers, we confirmed that these tags, if present, were typically available within the audience-facing HTML of an online news article.

News Publisher Sample Design

Given our interest in exploring the prevalence of platform-facing digital markup across news publishers serving a variety of news audiences, we curated an intentional sample that included large, "digital native", geographically and identity-focused news organizations in the United States by leveraging a range of existing news publisher lists. For example, we used the Pew Research Center's 2020 State of the News Media fact sheets (*State of the News Media methodology* 2020) to identify "large" (based on Sunday print circulation) and "online" (10 million or more monthly visitors) news organizations. For local news organizations, we turned to FreedomForum.org's "Today's Front Pages" website to identify 365 local news organizations with a digital presence. To identify news organizations targeting Black audiences, we leveraged the work of the Craig Newmark School of Journalism's Black Media Initiative, which includes a directory of approximately 350 Black-owned and focused news organizations (*Directory of Black Media Organizations* 2020). Finally, we leveraged resources at Harvard University (Tosat 2019), Baruch College and the Library of Congress (*LOC Ethnic Newspapers* 2022) to build a list of over 700 news organizations focused on serving both national audiences, as well as ethnic and diaspora communities. After refining these lists to eliminate overlaps, organizations that were no longer publishing or did not have an online presence, as well as those whose primary medium was broadcast (e.g., websites for radio stations), our sample of news organizations included:

- 50 "large" news organizations
- 37 "online-only" news organizations
- 326 "local" news organizations
- 184 news organizations focusing on Black audiences
- 145 news organizations focusing on ethnic and diaspora communities

Data Collection

We began our data collection with a pilot study to assess whether there was significant variation in the markup on publications' article pages according to the "section" (e.g., "business" versus "arts" etc.) in which it appeared, and describe this evaluation in section 3.3.1. As detailed in section 3.4, we are mindful of the potential negative impact that automated data collection practices may have on digital publishers, and therefore sought to minimize the volume of data downloaded from each publisher while still ensuring the validity and integrity of our results. Based on the results of this pilot study, we concluded that adapted our data collection process for the full sample, as described in section 3.3.2.

Pilot Study

The goal of the pilot study was to confirm whether the presence of Schema, Facebook, and Twitter tags on a given new organization's "article" pages was consistent across news sections (e.g., local/national, business/finance, arts, etc.), in order to minimize the number of articles downloaded from each publication, while maintaining

the accuracy of the metadata profile for each organization. After grouping the news organizations within each publisher category described in section 3.2 according to characteristics such as size, geography and publication frequency, we used a basic randomization function to select 10 news organizations per category, evenly distributed across these sub groupings. We then visited each publisher's site and manually selected the two most recent articles within each of five sections. Because organizations' coverage of news topics varied, for the pilot study we selected two articles from each of 5 sections in an effort to account for a breadth of topics within each organization. For each publisher this resulting in reviewing two articles from some combination of the following list of coverage areas.

1. Local/national/international
2. Opinion/editorial/analysis
3. Business/finance/community
4. Arts/culture/entertainment/food/lifestyle
5. Sports/real estate/transportation/science

A review of the Schema, OpenGraph, and Twitter markup/metadata tags present on the 500 article pages in our pilot sample (two article pages from each of five sections on the websites of 10 publishers chosen at random from our five publisher categories) revealed no meaningful difference in tag use across sections, with variations only arising when the primary content format of the article was e.g., video, audio or graphics. We therefore concluded that an article's "section" did not need to be factored into the design of our full study, which focused on text-based articles only. The design of the fully study is described in the next section.

Full Study

Given the consistency in the digital semantic infrastructure of Schema, Facebook (OpenGraph) and Twitter markup found across content verticals within each website in our pilot study, we proceeded to collect the HTML of three articles from the front pages of each of the 742 news organizations in our sample, for a total of 2,226 pages. To do this, we created a spreadsheet for each publisher category, and manually entered the full URLs of three articles selected from the front page during June, 2021. In an adjacent column, we generated a simple "shortcode" for each article that could be used to uniquely identify it. We then used these spreadsheets as input to a custom-built Python script that downloaded the contents of each URL and saved it locally under the shortcode identifier. We then constructed another set of Python scripts to parse the contents of the HTML files. Specifically, we created one script that analyzed all input files for the ld + json tag attribute and created a single column with all contents of that tag. We then used these initial results to create an inclusive dictionary of relevant attributes, and re-ran our script to separate the value of each attribute into a separate column in our results spreadsheet. We then repeated similar processes with additional Python scripts for the Facebook OpenGraph (via tags with the og: property) and Twitter tags (via tags with the twitter: property). In all, we captured the value (or

absence) of 24 unique Schema attributes, 16 Facebook/OpenGraph tag attributes, and 11 unique Twitter tag attributes per article page.

Ethics Considerations

We carefully considered the ethics of this study and made a number of data collection decisions based on ethical considerations. While this work does not involve research on human subjects and is therefore not subject to IRB review, there are still risks, benefits, and ethical considerations to be balanced in this research process.

Given our interest in exploring the potential patterns and differences in the way that news organizations targeting different audiences may be able to leverage the nominally "neutral" digital semantic infrastructure of platform-facing metadata structures, we took great care in curating our sample to include a broad range of news organizations—especially smaller, specialty, ethnic, and Black-serving publications. This is especially important given that "mainstream" journalistic organizations are not demographically reflective of the U.S. population (Merrefield 2020; Gray 2020).

Recognizing that many digital news publishers operate in constrained resource contexts, we also sought to minimize the costs imposed by our work on the news organizations we included in our study, principally by downloading only the minimum number of article pages required to ensure the validity of our results. This is why we began with a pilot study collecting 10 pages per publisher, which revealed sufficient markup consistency that we were confident a 3-article sample from each news organization was sufficient to profile their markup for the full study sample. Additionally, while we did use automated tools to download publishers' web content, we made no effort to circumvent any news organization's paywall or metered content limits. Where collecting the entire contents of a web page could not be accomplished through a simple automated download of the source code, we collected the content manually instead, with one of the researchers visiting the page directly—providing a valid email address or login if required—and saving a copy of the Web page manually. In neither the pilot study nor the full study did this process surpass the limits imposed on free access to any news organization's content. Finally, we ensured that our automated download script provided complete and valid contact information for the research team in the request header, which could be used to contact us in case our research process proved problematic for publishers at any time. To date, we have received no complaints from publishers regarding our data collection process or the demands it imposed on publisher systems.

Results and Data Analysis

In this section, we present the results of our survey of online news publishers' article web pages, with a focus on the presence and content of platform-facing markup tags. We then analyze these data in the context of publishers' format (e.g., online-only) and audience of focus, and extend our analysis to include crucial information about publishers' content management systems (CMSs), which our analysis has identified as a strong predictor of tags used in publishers' digital news articles.

Prevalence of Platform-Facing Metadata

Using custom Python scripts as noted in Section 3.3.2, our initial layer of analysis focused on top-level presence of specific tags associated with the Schema.org, Facebook, and Twitter metadata specifications: a < script type="application/ld + json"> tag, and < meta > tags containing property values that began with either og or twitter. We note that while the current Google-recommended format for Schema.org tags consists of a single < script type="application/ld + json"> containing a json-formatted blob encompassing a wide variety of possible attributes (Loosen 2002), both Facebook and Twitter metadata is designed as a series of discrete—if interdependent—array of < meta > tags with distinct properties. For this initial layer of analysis, we considered that a news organization supported a specific type of metadata if any appropriately formatted platform tag existed on at least one of the news article pages retrieved.

As shown in Table 1, the prevalence of support for platform-facing Schema.org, Facebook, and Twitter tags on news article pages roughly corresponds to the audiences served by publishers: nearly all large, online and regionally focused news organizations' article pages contain at least the top-level tags corresponding to each set of protocols, while the prevalence of these tags on Black and ethnically focused news organizations article pages is at most 80%, and in some cases less than 60%.

Robustness of Platform-Facing Metadata

Following this top-level analysis we sought to answer our RQ1 and understand the comparative *robustness* of this metadata across publishers by analyzing the range of attributes that a given publisher populated for each type of platform metadata. While our initial analysis provided insight into the fundamental *capacity* of various news organizations to include platform-facing metadata on their news articles, this second layer of analysis was designed to assess the degree to which this capacity was being leveraged to surface meaningful information about a news article to a given platform.

For Schema.org tags, we accomplished this by iteratively developing a dictionary of attributes found in *any* of the 2,226 individual pages in our corpus, eventually selecting 24 top-level attributes whose value (or absence) we tallied for each news article page across all publishers. While we included many attributes in this assessment that actually appeared on only a handful of article pages, our primary goal in this phase of the analysis was to understand the breadth of real-world tag use on article pages published by actual news organizations, given the more than a hundred valid attributes

Table 1. Number and percentage of organizations in each category that show potential support for each type of platform-facing markup tag.

		Basic support available for		
Organization category	No. of orgs.	Schema	Facebook	Twitter
Pew Big	50	48 (96%)	50 (100%)	49 (98%)
Pew Online	37	37 (100%)	37 (100%)	33 (89%)
Local	326	293 (90%)	319 (98%)	300 (92%)
Black	184	107 (58%)	143 (79%)	121 (65%)
Ethnic	145	84 (58%)	114 (78%)	94 (66%)

for e.g., the NewsArticle content type under the Schema.org specification (*NewsArticle: A Schema.org Type* 2022).

For both the Facebook og and Twitter twitter tags, analysis was somewhat more straightforward. In part, this is due to the discrete and limited nature of the metadata implementation for these platforms, in which each attribute is encoded via an individual tag, and there is no nesting of values[1]. By contrast, Facebook's metadata specification has only four required attributes (title, type, image and url), seven general-purpose optional attributes and six more optional attributes associated with the article type (*The Open Graph protocol* 2022). Twitter, meanwhile, enumerates 22 distinct Twitter "Card" attributes; but more than half of these are specific to either the Twitter "player" card—which is intended specifically for non-text content—or the Twitter "app" card, which is relevant specifically to mobile apps (*Twitter Cards Guide* 2021).

As illustrated in Table 2, analyzing the *robustness* of each news organization's platform-facing markup yields a somewhat more complex picture of the possible interaction between each outlet's top-level classification and the degree to which it engages this type of digital infrastructure.

Taken together, these results illustrate how asymmetrical robust use of platform-facing metadata truly is: while 42/50 (84%) of large and 33/37 (89%) of online-only news organizations make robust use of Schema.org tags, only 219/326 (67%) of local news organizations, 39/145 (27%) of ethnic, and 20/184 (11%) of Black-focused news organizations do the same. These findings highlight the fact that there is a relatively weak relationship between technical *support* for platform-facing tags and robust *use* of those tags, especially among local news organizations.

These discrepancies therefore led us to develop a third and final research question, which we were able to address through an additional layer of analysis, with a focus on the technologies that drive the online-publishing functionality of every digital newsroom: content management systems (CMSs):

> RQ3: What correlations, if any, exist between a news organization's CMS and the characteristics of the markup included on its text-article pages?

CMS Prevalence

Because CMSs may influence the markup found on published news article web pages, we began this part of our analysis by conducting a multi-layered review that included

Table 2. "Weak" support is defined in comparison with the overall sample: For Schema, it means at most the @context, type and url attributes have values; for Facebook, it means there are six or fewer attributes; for Twitter it means five or fewer attributes.

		Weak attribute completion for		
Organization category	No. of orgs.	Schema	Facebook	Twitter
Large	50	8 (16%)	9 (18%)	18 (36%)
Online-only	37	4 (11%)	11 (30%)	16 (43%)
Local	326	107 (33%)	39 (12%)	143 (44%)
Black	184	87 (47%)	36 (20%)	108 (59%)
Ethnic	145	45 (31%)	24 (17%)	75 (52%)

Note: This table shows only those organizations that could support tags, but have weak attribute completion.

the source code on article and Terms of Service pages, as well as press releases and other published material, in order to identify the CMS used by each publisher. Because the HTML markup of a news article page often contains identifying elements that refer to the CMS or site builder—such as WordPress or BLOX—reviewing this markup (which we had already downloaded as part of our tag review) often yielded definitive indicators for these CMSs. Additionally, likewise, the URL from which photos or data files for a website widget originate can also offer clues about the specific CMS used for publishing. In several cases, publishers actually listed the name of software-as-a-service (SaaS) tools in the footer of their websites or on the Terms of Service page. Finally, press releases and public organizational memos often recounted partnership contracts and other information regarding a publisher's use of a particular CMS. Using this approach, we were able to identify the CMSs used by 685 out of 742 publications in our corpus (92%). We display the CMSs used by a minimum of five publications in our corpus and the total number of publications that use the listed CMS in Table 3. WordPress accounts for a quarter of all CMSs in our sample, and the top eight CMSs (WordPress, BLOX, Presto," Unknown," WordPress VIP, McClatchy, WP Bakery, and ARC) account for three-quarters of all CMSs in our sample.

Finally, to provide further context for our analysis, we reviewed published information about the availability and cost of the most popular CMSs within our sample. In the next section, we present these findings alongside our prevalence and other results.

Table 3 illustrates that while more than 50% of all news organizations in our sample rely upon WordPress, BLOX or Presto, the institutional and financial availability of these CMSs varies dramatically. While WordPress is free and open-source, we estimate that BLOX costs upwards of $1,500 per month and Presto is fully proprietary and exclusive to Gannett publications.

Table 3. Prevalence of CMS across our full sample of 742 news organizations.

CMS name	No. of orgs	Pct. of sample	Estimated cost (USD)
WordPress*	188	26%	0-59
BLOX	127	17%	>1500
Presto	113	15%	Proprietary
Unknown	47	6%	
WordPress VIP	30	4%	<2000
McClatchy**	24	3%	Proprietary
WP Bakery	20	3%	***
ARC	15	2%	<10,000
Advance Local Media **	10	1%	Proprietary
CUE	8	1%	
WordPress / Google Site Kit	8	1%	0
Hearst**	7	1%	Proprietary
Wix	7	1%	0-500+
Chorus	6	1%	<8,000
Drupal	6	1%	<500
AMP	5	1%	0
Joomla!	5	1%	0
NewsEngin Ampere	5	1%	
Presteligence/MyNews360	5	1%	
WordPress / Visual Composer	5	1%	4-71

Estimated Cost is rounded to the most accurate US dollar possible, and averaged per month. 'Proprietary' label refers to CMSes not for sale. Missing values are unknown. *All WordPress versions (e.g., 5.8.1) were grouped under an umbrella WordPress category, with the exception of those that were distinctly named/branded. **Specific CMS name unknown. *** Lifetime Assess only offered, with cost ranging from 56-299 USD.

When we further compare the prevalence of certain CMSs across our five publisher categories, moreover, we find additional patterns of interest:

- Among **large** news organizations (n = 50), one-third (34%) are using a proprietary CMS that is not for sale; roughly another third (34%) use a CMS (specifically ARC, WordPress VIP and BLOX) that costs a minimum of $1000 per month.
- Among **online-only** news organizations (n = 37), the choice of CMS is far more variable (only Chorus use, at 14%, claims a double-digit percentage of publishers). Apart from WordPress VIP, there is no overlap between the most common CMSs in this group and those that are popular across our sample as a whole.
- Among Black news organizations (n = 184), 70% of the organizations in our sample use some version of WordPress (generic = 64%; WP Bakery = 6%). BLOX accounts for 4% of CMSs in this group.
- While the number of unidentified CMSs within our sample of ethnic news organizations (n = 145) was higher (24 orgs/17%), the three most popular CMSs we were able to identify were the same as among Black news organizations. WordPress accounts for more than half (75 orgs/52%) of all CMSs, BLOX for 9% (13 orgs) and WP Bakery for 6% (9 orgs).
- Among local news organizations (n = 326), we see significant use of proprietary and/or relatively expensive CMSs: almost three-quarters (73%) use a CMS that is either proprietary (Presto/111 orgs/30%, McClatchy/24 orgs/7%) or costs a minimum of $1,000 per month (BLOX/107 orgs/29%, WordPress VIP/26 orgs/7%). This stands in particular contrast to the CMSs used by Black and ethnic news organizations.

In the next section, we assess the relationship between CMS selection and the presence of specific metadata attributes within the Schema, Facebook, and Twitter tags.

Prevalence of Metadata Tags by CMS

Having successfully identified the CMSs used by the majority of the organizations in our sample, our next step was to respond to our RQ3 and compare the specific Schema, Facebook, and Twitter tags that were present on the pages of different publishers using a given CMS. To do this, we designed an R script that compared the tags present on web pages across all publishers using a given CMS. To better understand which tags were *likely completion* by a given CMSs' default page configuration, we then assessed the degree of overlap between, for example, Schema tags appearing on the news article pages of all news organizations using a given CMS. Starting with a 95% rate of overlap and moving down in 5% increments to 50% rate of overlap, we identified the Schema, Facebook, and Twitter tags used on 80% of all article pages produced by a given CMS.

Our analysis revealed that for the eight most popular CMSs, 6 had no change in the Schema tags used below the 80% threshold—suggesting that the Schema tags used by those 80% of publishers are strongly influenced by the default configuration of the CMS itself. We therefore define the sets of Schema, Facebook, and Twitter

metadata tags, respectively, used by at least 80% of publishers using a particular CMS as the "canonical" tag set for that CMS. Detailed results for these tag sets can be found in Tables 4, 5, and 6.

As illustrated in 4, the prevalence of Schema tags varies considerably by CMS, with expensive (>$2 K/month - i.e., Wordpress VIP, ARC, Chorus and CUE) and/or proprietary CMSs (i.e., Presto, McClatchy, Advance Local Media and Hearst) typically including 11–17 Schema tags on each online article page.

By contrast, the tags present on the pages generated by popular, but less expensive (i.e., BLOX, WP Bakery) or free (i.e., WordPress variations, Drupal, AMP), CMSs indicate support for Schema through the existence of an ld + json tag that contains some variation of the value schema.org in its @context tag. Apart from this, however, most of the pages in our sample generated by these CMSs contained few, if any, additional tags: BLOX appears to populate the @type and url tags, but the others contain no other tag values at the 80% threshold. The notable exception to this is Wix, which publishers in our sample were populating with 10 additional tags, including publisher, headline, datepublished, datemodified, url, and description. At the same time, Wix-driven publications account for only 1% of those in our overall sample, while WordPress and BLOX combined are used by 43% of the organizations in our overall sample, as well as 74% of Black publications and 67% of ethnic publications. Among local news organizations, however, these account for only 32% of publications, 8% of "large" publications, according to Pew's list. Also noteworthy is that *none* of Pew's "online-only" news organizations use either generic/low cost WordPress or BLOX.

Examining Tables 5 and 6 shows that these highly asymmetric patterns of tag population do not hold for either Facebook OpenGraph tags or for Twitter tags for popular CMSs in our sample. As Table 5 indicates, for example, OpenGraph tag usage is fairly consistent across *all* CMSs—with many free and low-cost systems typically populating even values like locale, whereas more expensive/proprietary systems do not. In fact, with the exception of WordPress Visual Composer, a one-time-fee plugin for WordPress, the only OpenGraph tags that are consistently (though not exclusively) populated by more expensive/proprietary systems are those related to image dimensions. Meanwhile, Chorus (the pricey, but popular with-online-only publications CMS) is alone in populating the alt-text of its OpenGraph images, and only BLOX contains section information. With respect to Twitter tags, as shown in 6, the field is somewhat more mixed, with similar trends in tag population to what was found in Schema tags. Once again, proprietary and more expensive systems populate more tags, often including description and title information left off of less expensive CMS pages. Interestingly, however, Presteligence and NewsEngin—neither of which had Schema tags in our canonical sample—each include multiple Twitter tags on their pages. Also noteworthy is the fact that CUE does not appear to include any Twitter tags at all.

Discussion

This study began by examining the markup embedded in digital news articles (RQ1) and how it varied—if at all—based on news organizations' characteristics (RQ2). Our analysis subsequently revealed the importance of specific types of content

Table 4. Canonical Schema tags across all news orgs using the indicated CMS.

	Word Press	BLOX	Presto	Wordp Press VIP	McClatchy-Unknown	WP Bakery	ARC	ALM-Unknown	WordPress/ Google Site Kit	CUE	WIX	Hearst-Unknown	Drupal 7	Chorus	WordPress Visual Composer	AMP	Prestiligence/ MyNews360	NewsEngin Ampere	Joomla!
No. of orgs	188	127	113	30	24	20	15	10	8	8	7	7	6	6	5	5	5	5	5
No of tags.	1	3	17	15	17	1	13	15	1	11	10	11	1	14	1	1	0	0	0
Supports Schema?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X	X
keywords			•	•	•			•						•					
url		•	•	•	•		•	•			•			•					
publisher			•	•	•		•	•		•	•	•		•					
@type		•	•	•	•		•	•		•	•	•		•					
mainentityofpage			•	•	•		•	•		•	•	•		•					
description			•		•		•			•	•	•		•					
datemodified			•	•	•		•	•		•	•	•		•					
datepublished			•	•	•		•	•		•	•	•		•					
image			•	•	•		•	•		•	•	•		•					
author			•	•	•		•	•		•		•		•					
headline			•	•	•		•	•		•	•	•		•					
haspart			•				•			•									
isaccessibleforfree																			
creator				•	•														
datecreated			•	•															
dateline																			
thumbnailurl			•	•	•									•					
articlesection			•	•	•									•					
ispartof			•		•		•	•											
inlanguage																			
articlebody												•							

Table 5. Canonical Facebook tags across all news orgs using the indicated CMS.

	Word Press	BLOX	Presto	Wordp Press VIP	McClatchy-Unknown	WP Bakery	ARC	ALM-Unknown	WordPress/ Google Site Kit	CUE	WIX	Hearst-Unknown	Drupal 7	Chorus	WordPress Visual Composer	AMP	Prestiligence/ MyNews360	News Engin Ampere	Joomla!
No. of orgs	188	127	113	30	24	20	15	10	8	8	7	7	6	6	5	5	5	5	5
og:url	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
og:description	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
og:type	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
og:site_name	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	
og:locale	•			•		•	•		•			•			•	•	•		
og:title	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
og:image	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
og:image:width		•	•	•		•		•	•	•	•			•	•	•			
og:image:height		•	•	•		•			•	•	•			•	•	•			
og:section		•																	
og:image:type																			
og:image:secure_url																			
og:pixelID																			
og:image:alt														•					
g:image:url																			

Table 6. Canonical Twitter tags across all news orgs using the indicated CMS.

	Word Press	BLOX	Presto	Wordp Press VIP	McClatchy-Unknown	WP Bakery	ARC	ALM-Unknown	WordPress/ Google Site Kit	CUE	WIX	Hearst-Unknown	Drupal 7	Chorus	WordPress Visual Composer	AMP	Prestiligence/ MyNews360	News Engin Ampere	Joomla!
No. of orgs	188	127	113	30	24	20	15	10	8	8	7	7	6	6	5	5	5	5	5
Twitter:card	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•
Twitter:site		•	•		•		•	•				•		•			•	•	
Twitter:image			•	•	•		•	•			•	•		•		•	•	•	
Twitter:title			•		•		•	•			•	•	•	•		•	•	•	
Twitter:creator			•				•		•										
Twitter:url													•	•					
Twitter:description			•		•		•	•			•	•		•		•	•	•	
Twitter:image:alt		•			•									•					
Twitter:dnt												•							
Twitter:domain																			

management systems (CMSs) with respect to metadata robustness, leading to our final research question (RQ3), which explored the way that CMS availability correlates with both the organization's financial resources and institutional relationships. In this section, we unpack the implications of these results with respect to the discoverability of online news articles produced for and by certain communities, and how this does or does not reconcile with the conceptual "neutrality" of the digital semantic infrastructure for online publishing (Kennedy and Griffith 2020; Christin 2020; Bowker et al. 2009) that these markup and metadata protocols represent.

The digital infrastructure identified in this study represents the metadata-driven interface between platforms that rank, curate, and distribute journalistic content, and the publishers that produce them. While previous scholarship has identified modes of connections between platforms and news organizations (Ananny 2018), and discussed the political-economic forces shaping the institutional and business relationship between news organizations and technology companies (De Maeyer 2019; Hardy 2017), no deep analytical attention has been given to the digital semantic infrastructure upon which the automated curation and distribution of digital news stories depends. In the following sections, we discuss the implications of our findings for understanding the role and impact of CMSs and markup practices as foundational components of digital publishing infrastructures. We provide recommendations for both researchers and publishers.

Intersection of Content Management Systems and Markup

By investigating the markup patterns of digital news stories, this study revealed deep disparities in news organizations' ability to make use of the digital semantic infrastructure of platform-facing metadata structures; yet, these fault lines do not reflect a neat pattern of legacy print versus newer online media outlets or even (quite) mainstream versus marginal perspectives. Instead, we find that news organizations' "choice" of content management system (CMS) is the most reliable indicator of the robustness of their articles' markup, aligning with prior work that identifies these systems as important supporting technologies for generating quality digital journalism (Ojo and Heravi 2018).

We scare-quote "choice" in this instance precisely because the CMSs associated with the most robust metadata in our analysis are not available on an open market. To take one example, our results show considerable discrepancies in how Schema.org tags are deployed on pages produced by free and/or open-source CMSs, software-as-a-service (SaaS) website builder platforms, and bespoke CMSs available only to affiliates of a particular media conglomerate (i.e., McClatchy or Gannett). In the first two instances, both open-source WordPress pages and SaaS BLOX pages consistently included the foundational Schema.org ld + json tag, but published few or no meaningful attributes *within* it—indicating that while these CMSs *nominally support* e.g., the Schema.org markup standard, they do not actually populate them with useful content.

This stands in stark contrast to the markup found on pages produced by bespoke CMSs, such as those used by McClatchy, Gannett, or Advance Local Media affiliates. On these pages, the typical ld + json tag was populated with more than 15 unique attributes that meaningfully express the nature of the page's content in the platform's

preferred format. These attributes include clearly structured keyword and publisher information, a short description of the content and the date it was published, as well as the author and headline information, among other details. As a result, the search engine visibility (*Understand how structured data works* 2021) of the content on these pages is much greater.

We note in particular that WordPress sites—which are overwhelming implemented by Black and ethnic news organizations in our sample—include far fewer attributes than even relatively small news organizations that are affiliated with a larger media conglomerate. At the same time, we note some key exceptions to this broader pattern. For example, while the BLOX publishing system is specifically designed for digital news publications, it populates only the url and @type Schema tags, while the broadly targeted and relatively low-cost Wix platform populated nine additional Schema tags, suggesting that platform's greater focus on search-engine optimization (SEO) in the default design of their system.

By contrast, Facebook OpenGraph tags were used far more uniformly than Schema.org tags across CMSs. For example, all of the top CMSs in our survey populated at least six Facebook OpenGraph tags—many of which parallel tags included in the Schema.org specification but were *not* populated by the same CMS. Thus while our canonical WordPress OpenGraph tag set included url, description, and type, the seemingly parallel Schema.org tags (i.e., 'url, description, @type) were *not* present on the majority of WordPress pages in our sample, indicating that either CMS developers or publishers have specifically prioritized populating these social media-focused tags.

While prior research has chronicled platforms'—especially Facebook's—shifting strategies toward news partners and digital news content (Bucher 2021), and further research is needed regarding the particular technological decisions behind the design of OpenGraph's syntax, we suggest a few possible reasons for the discrepancy between OpenGraph and Schema.org tags based on their co-occurrence patterns found in our sample. First, we note that Twitter tags show a broadly similar pattern of usage to what we observe with respect to Schema.org tags: proprietary and/or pricier CMSs include more Twitter tags than their less expensive/open source counterparts. Notably, however, in the absence of Twitter-specific metadata tags, the social media platform will fall back on information in OpenGraph tags to generate Twitter "Cards" for content shared on the platform. Thus, it may be that the lack of detail in Twitter tags reflects an expectation that OpenGraph tags will do "double duty" in providing sufficient metadata for social media sharing and discovery of news publishers' content:

> Facebook's directing that kind of traffic because it wants to direct that traffic—it wants to be a digital publishing kingmaker … [By contrast] SEO journalism just doesn't make commercial sense anymore. Social trumps search, at least when it comes to the attention that sells ads. (Meyer 2014)

This does suggest that Facebook's active promotion of its metadata tagging system, combined with a concerted effort to drive revenue-producing traffic to news publishers, may have had a lasting impact on this aspect of digital news infrastructure. While our results are far from definitive, we note that this difference in usage is *not* explained by a difference in longevity, as Schema.org tags were introduced in 2011, just one year after Facebook's OpenGraph tagging system.

Digital Publishing: Open Field or Stacked Deck?

Our findings suggest that the identified discrepancies in markup practices and CMSs available to different news publishers may reflect and exacerbate existing inequities in journalistic representation—particularly through creating barriers to the visibility of content produced by news organizations serving Black and ethnic communities. The underrepresentation of certain voices and perspectives in U.S. journalism is well-documented (Sherry and Matsaganis 2019; Gray 2020; Belair-Gagnon, Nelson, and Lewis 2019). In the print era, Black and ethnic media organizations, in particular, suffered from chronic underinvestment and sometimes outright attack (Walker 2020). Yet while the shift to digital publishing that has characterized the past 20 years has destabilized the print journalism industry—with some even suggesting that the key to a successful transformation should start with bankruptcy (Langeveld 2009)—the comparatively low overhead of starting digital ventures initially led to astronomical growth of online-first newsrooms (Bump 2014). While this growth in online publications has included many Black, Hispanic, ethnic and diaspora-focused news organizations, they have continued to face special challenges. This includes difficulty securing advertising dollars because their target audiences are undervalued, and because they may be excluded from circulation assessments that would allow them to command higher prices for advertisements (Abernathy 2020).

These newer, community-focused digital journalism ventures also tend to be independent (Abernathy 2020). In addition to the typical business challenges that this creates, they also cannot publish via the expensive or bespoke CMSs that our results indicate produce the most robust platform-facing metadata; as a result, they may face a vicious cycle of reduced online visibility and circumscribed audience growth. Without sufficiently descriptive metadata, these outlets are less likely to be well-indexed by the very platforms—such as Google and Facebook—that audiences not only rely on to discover information, but which also control the vast majority of online ad revenues. In the next section, we discuss future directions for research to further understandings of potential barriers to representation posed by CMSs and markup practices and how they could be navigated. We also provide recommendations for best practices for digital news publishers.

Investing in Infrastructure: Improving the Markup Practices of Digital News

With future research areas in mind, our analysis prompts the question of whether enhanced technical training in web development and digital publishing tools could improve digital publishers' markup practices, with a special focus on the technical-administrative work of editors and developers who work for Black or ethnic media organizations. These research questions may be proposed within the context of the broader institutional analysis of the costs, affordances, and interoperability characteristics of the various CMSs. Do off-the-shelf website-builder platforms allow markup editing, and if so, under what permission regime? Do CMS costs positively correlate with increased levels of sophistication with which indexing and markup practices can be carried out. In short, is infrastructure-accessibility a resource limitation issue? For example, the WordPress platform supports an ecosystem of plugins that could

potentially ease the integration of more robust markup into publishers' content, but these solutions may be brittle, undersupported, or overly technical for many smaller organizations to leverage. This indicates that deeper investigation into the actual affordances of widely used CMSs as well as ethnographic research regarding newsroom resources and practices is needed to understand what types of interventions or innovations are likely to result in more robust metadata annotations and, in turn, better representation for publishers in the general platform economy of journalism.

Future research into the digital infrastructure maintaining the interface between news organizations and platforms is critical. Infrastructures exert their power through their hidden affordances with which they encourage or limit accessibility to their services. In our research we have theorized that markup and metadata are the machine-readable interface that determines a digital news content's rank, classification, and other properties in a list of search results or cascading content feeds, and thus plays an invisible but significant role in determining the discoverability and representation of particular news stories. In short, an underdeveloped markup layer with insufficient metadata annotations can jeopardize the indexing and rendering of a story page in news feeds. Weak metadata can also cause further problems, such as lack of attribution, incorrect geotagging, or rendering the incorrect image for a news story once a news aggregator begins hosting it. For smaller news organizations with fewer resources to address this technological component of their publishing process, our research suggests that it may prove beneficial to establish a practice of meticulously and consistently marking up their digitally published work.

Our findings have implications for research that probes how algorithmic curation infrastructures shape news content, visibility, and representation (Diakopoulos 2015; Diakopoulos 2020; Sherry 2015; Sherry and Matsaganis 2019) by identifying CMSs and the associated markup patterns as important factors that may influence the discoverability of digital news. While some work has begun to explore relationships between markup quality and information quality in identifying "fake news" items (Kennedy and Griffith 2020; Castelo, et al. 2019), future work should further unpack the relationships between the CMS used by a news publisher, the semantic infrastructure of their digital news pages, and differences in how news content is surfaced by various platforms' algorithms. Particular attention should be paid to the implications that CMS choice and markup practices have for traditionally underrepresented voices.

Conclusion

In sum, we find the semantic infrastructure embodied in the markup layer of digital news stories varying significantly in the implementation of—if not support for—Schema.org, Facebook/Open Graph, and Twitter metadata tags. In addition, however, we find a possible association between metadata quality and broad technological-institutional asymmetries in the form of the content management systems used by news organizations that reflect the expertise—as well as the expenses—associated with those CMSs. In other words, publishing patterns found in the semantic infrastructure of platform journalism appear to be closely associated with the technological affordances of the content builder platforms that underpin the markup layer. These

findings highlight the extent to which—despite the "free" and "open" nature of these markup specifications—building a robust digital infrastructure of platform visibility for journalism using these theoretically accessible specifications still depends on news organizations' access to relatively traditional resources and relationships.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1940670, 1940713 and 1940679.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Note

1. In Schema.org, for example, the value of an author or publisher attribute may itself be an object with layered content, such as @type, name, url etc.

ORCID

Emily Sidnam-Mauch http://orcid.org/0000-0003-3372-7778

References

Abernathy, P. M. 2020. The Expanding News Desert. *The Center for Innovation and Sustainability in Local Media, UNC Chapel Hill*. https://www.usnewsdeserts.com/reports/news-deserts-and-ghost-newspapers-will-local-news-survive/the-news-landscape-of-the-future-transformed-and-renewed/journalisticmission-the-challenges-and-opportunities-for-ethnic-media/

Ananny, M. 2018. *Networked Press Freedom: Creating Infrastructures for a Public Right to Hear*. Cambridge, MA: MIT Press.

Ananny, M., and K. Crawford. 2015. "A Liminal Press: Situating News App Designers within a Field of Networked News Production." *Digital Journalism* 3 (2): 192–208.

Bandy, J., and N. Diakopoulos. 2020. "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News." In:*Proceedings of the International AAAI Conference on Web and Social Media* 14: 36–47.

Belair-Gagnon, V., J. L. Nelson, and S. C. Lewis. 2019. "Audience Engagement, Reciprocity, and the Pursuit of Community Connectedness in Public Media Journalism." *Journalism Practice* 13 (5): 558–575.

Bowker, G. C., and S. L. Star. 1998. "Building Information Infrastructures for Social Worlds— the Role of Classifications and Standards.". In: *Community Computing and Support Systems*, 231–248. Berlin: Springer.

Bowker, G. C., K. Baker, F. Millerand, and D. Ribes. 2009. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In Hunsinger, J., Klastrup, L., Allen, M. (eds) *International Handbook of Internet Research*. Springer, Dordrecht.

Bucher, T. 2017. "The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms." *Information, Communication & Society* 20 (1): 30–44.

Bucher, T. 2021. *Facebook*. Polity Press, Medford.

Bump, P. 2014. Starting a News Organization? Here's How You'll Make Money. *The Atlantic*. https://www.theatlantic.com/national/archive/2014/03/startinga-news-organization-heres-how-youll-make-money/359662/

Burgess, J., and E. Hurcombe. 2019. "Digital Journalism as Symptom, Response, and Agent of Change in the Platformed Media Environment." *Digital Journalism* 7 (3): 359–367.

Castelo, S, et al. 2019. "A Topic-Agnostic Approach for Identifying Fake News Pages." In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 975–980.

Christin, A. 2020. "What Data Can Do: A Typology of Mechanisms." *International Journal of Communication* 14: 1115–1134.

De Maeyer, J. 2019. Content Management Systems and Journalism. In: *Oxford Research Encyclopedia of Communication*. Published online by Oxford University Press. Accessed 5 September 2021.

Diakopoulos, N. 2015. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures." *Digital Journalism* 3 (3): 398–415.

Diakopoulos, N. 2020. "Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism." *Digital Journalism* 8 (7): 945–967.

Directory of Black Media Organizations 2020. Newmark Journalism School Center for Community Media. https://airtable.com/shrKbdiGOaRdsSIIW/tblPDC9g46NM1n7Np

Dowd, C. 2020. *Digital Journalism, Drones, and Automation: The Language and Abstractions behind the News*. Oxford: Oxford University Press.

Erwig, M. 2017. *Once upon an Algorithm: how Stories Explain Computing*. Cambridge, MA: MIT Press.

Fensel, D., F. van Harmelen, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider. 2001. "OIL: An Ontology Infrastructure for the Semantic Web." *IEEE Intelligent Systems* 16 (2): 38–45.

Gray, K. 2020. The racial divide on news coverage, and why representation matters. *Knight Foundation*. https://knightfoundation.org/articles/the-racial-divideon-news-coverage-and-why-representation-matters/

Guha, R. V., D. Brickley, and S. Macbeth. 2016. "Schema. org: Evolution of Structured Data on the Web." *Communications of the ACM* 59 (2): 44–51.

General structured data guidelines (2022. Google Search Central. url:https://developers.google.com/search/docs/advanced/structured-data/sd-policies

Hardy, J. 2017. "Money,(Co) Production and Power: The Contribution of Critical Political Economy to Digital Journalism Studies." *Digital Journalism* 5 (1): 1–25.

Hitzler, P. 2021. "A Review of the Semantic Web Field." *Communications of the ACM* 64 (2): 76–83.

IPTC NewsML 1.0 Specification 2002. https://www.iptc.org/std/NewsML/1.0/specification/NewsML_1.0-spec-functionalspec_4.html

IPTC NewsML G2.0 Specification 2021. https://iptc.org/standards/newsmlg2/whos-using-newsml-g2/

Kammer, A. 2021. "Resource Exchanges between Mobile News Apps and Third-Parties." *Digital Journalism*. Advance online publication. doi: 10.1080/21670811.2021.2000455.

Kennedy, C., and J. Griffith. 2020. "Using Markup Language to Differentiate between Reliable and Unreliable News." In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 619–625.

Kodama, M, et al. 2008. "Realizing a News Value Markup Language for News Management Systems Using newsML." In 2008 International Conference on Complex, Intelligent and Software Intensive Systems. IEEE, pp. 249–255.

Langeveld, M. 2009. Could strategic bankruptcies be needed to transform newspapers? *Nieman Lab*. https://www.niemanlab.org/2009/06/could-strategicbankruptcies-be-needed-to-transform-us-newspaper-enterprises/

LOC Ethnic Newspapers 2022. Library of Congress. https://www.loc.gov/search/? all=true&fa=subject:ethnic+newspapers

Loosen, W. 2002. "The Second-Level Digital Divide of the Web and Its Impact on Journalism." *First Monday* 7 (8). doi: 10.5210/fm.v7i8.977.

Merrefield, C. 2020. Race and the newsroom: What seven research studies say. *Nieman Lab*. https://www.niemanlab.org/2020/07/race-and-the-newsroom-whatseven-research-studies-say/.

Meyer, R. 2014. And Just Like That, Facebook Became the Most Important Entity in Web Journalism. *The Atlantic*. https://www.theatlantic.com/technology/archive/2014/02/and-just-like-that-facebook-became-the-most-important-entity-inweb-journalism/283618/

Mika, P. 2015. "On Schema. org and Why It Matters for the Web." *IEEE Internet Computing* 19 (4): 52–55.

NewsArticle: A Schema.org Type 2022. https://schema.org/NewsArticle

Ojo, A., and B. Heravi. 2018. "Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences." *Digital Journalism* 6 (6): 693–718.

OpenGraph Markup Guide 2021. https://developers.facebook.com/docs/sharing/webmasters#markup

Recordon, D. 2010. *The Open Graph Protocol*. https://www.w3.org/2010/04/w3c-track.html

Schema.org History of Development 2015. https://schema.org/docs/about.html

Schjøtt Hansen, A., and J. M. Hartley. 2021. "Designing What's News: An Ethnography of a Personalization Algorithm and the Data-Driven (Re) Assembling of the News." *Digital Journalism*. Advance online publication. doi: 10.1080/21670811.2021.1988861.

Schwartz, D. G. 2003. "From Open is Semantics to the Semantic Web: The Road Ahead." *IEEE Intelligent Systems* 18 (3): 52–58.

Sherry, S. Y. 2015. "The Inevitably Dialectic Nature of Ethnic Media." *Global Media Journal* 8 (2): 133.

Sherry, S. Y., and M. D. Matsaganis. 2019. *Ethnic Media in the Digital Age*. New York: Routledge.

State of the News Media methodology 2020. Pew Research Center. https://www.pewresearch.org/journalism/2021/07/27/state-of-the-news-media-methodology/

Sunne, S. 2021. "An Introduction to Schemas for Journalists." *Nieman Foundation*

The Open Graph protocol 2022. https://ogp.me/#types

Thorson, Kjerstin, Mel Medeiros, Kelley Cotter, Yingying Chen, Kourtnie Rodgers, Arram Bae, Sevgi Baykaldi, et al. 2020. "Platform Civics: Facebook in the Local Information Infrastructure." *Digital Journalism* 8 (10): 1231–1257.

Tosat, C. G. 2019. *Hispanic Digital Newspapers in the U.S., 2019: Evolution, quality, and impact*. https://cervantesobservatorio.fas.harvard.edu/en/reports/hispanic-digital-newspapersus-2019-evolution-quality-and-impact

Twitter Cards Guide 2021. https://developer.twitter.com/en/docs/twitterfor-websites/cards/overview/markup

Understand how structured data works 2021. https://developers.google.com/search/docs/advanced/structured-data/intro-structured-data

Van Dijck, J., T. Poell, and M. De Waal. 2018. *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press.

Walker, M. 2020. Honoring African American Contributions: The Newspapers. *Library of Congress Blog*. https://blogs.loc.gov/headlinesandheroes/2020/07/honoring-african-american-contributions-the-newspapers/