The News Crawler: A Big Data Approach to Local Information Ecosystems

Khanom, Asma; Kiesow, Damon; Zdun, Matt; Shyu, Chi-Ren

cogitatio

Media and Communication (ISSN: 2183–2439)
2023, Volume 11, Issue 3, Pages 318–329
https://doi.org/10.17645/mac.v11i3.6789

Article

The News Crawler: A Big Data Approach to Local Information Ecosystems

Asma Khanom [1,*], Damon Kiesow [1], Matt Zdun [2], and Chi-Ren Shyu [2]

[1] School of Journalism, University of Missouri–Columbia, USA
[2] Institute for Data Science and Informatics, University of Missouri–Columbia, USA

* Corresponding author (asma.khanom@mail.missouri.edu)

Submitted: 31 January 2023 | Accepted: 20 June 2023 | Published: 28 September 2023

Abstract

In the past 20 years, Silicon Valley's platforms and opaque algorithms have increasingly influenced civic discourse, helping Facebook, Twitter, and others extract and consolidate the revenues generated. That trend has reduced the profitability of local news organizations, but not the importance of locally created news reporting in residents' day-to-day lives. The disruption of the economics and distribution of news has reduced, scattered, and diversified local news sources (digital-first newspapers, digital-only newsrooms, and television and radio broadcasters publishing online), making it difficult to inventory and understand the information health of communities, individually and in aggregate. Analysis of this national trend is often based on the geolocation of known news outlets as a proxy for community coverage. This measure does not accurately estimate the quality, scale, or diversity of topics provided to the community. This project is developing a scalable, semi-automated approach to describe digital news content along journalism-quality-focused standards. We propose identifying representative corpora and applying machine learning and natural language processing to estimate the extent to which news articles engage in multiple journalistic dimensions, including geographic relevancy, critical information needs, and equity of coverage.

Keywords

critical information needs; information ecosystem; local news; machine learning; news deserts; United States

Issue

This article is part of the issue "News Deserts: Places and Spaces Without News" edited by Agnes Gulyas (Canterbury Christ Church University), Joy Jenkins (University of Missouri), and Annika Bergström (University of Gothenburg).

1. Introduction

In the past 20 years, the newspaper industry in the US has undergone immense disruption, from the digital revolution to the Great Recession, reversing the good fortunes of previous decades (Abernathy, 2020; Ali et al., 2020). The Covid-19 pandemic has further accelerated the crisis and expanded the news deserts in the US (Gabbatt, 2020). Many news outlets shuttered suddenly (Ferrucci, 2019; Griffin, 2018), leaving local readers in the dark.

To define "news desert," this study adopts Abernathy's (2018, 2022) study, which reported that growing areas (cities, towns, and regions) across the US are losing access to local news coverage. The decline in local coverage does not reduce the need for local information in local communities' day-to-day lives, and it harms residents (Hayes & Lawless, 2018). However, as news splinters and scatters, local readers have to actively seek information through newspapers in other towns or counties, digital-first news sites, television and radio broadcasters, and social media (Neff et al., 2022).

Some research into these changes has relied on a single-point geospatial location of a news outlet as a proxy for coverage area (Hutchins, 2022). This method effectively reveals national trends and informs discussion and policymaking (Jordon, 2018) but is an incomplete local measure, lacking an accurate estimation of newsroom contributions to its community's information needs, individually or in aggregate.

Alternatively, local news and information ecosystem audits have been performed. We are concerned primarily with "local news," which we find at the intersection of "the physical locations where reporting happens, where news–decision making occurs" (Usher, 2019, p. 86) and where the social and civic life of residents is centered. It is an inherently place-based but subjective categorization that discounts contributions to local information needs by non-journalistic sources.

The manual method of news content analysis deploys a team of researchers to collect news coverage from across a city or region for a period from just one day to even a week or more to evaluate the breadth and depth of the reporting, often using the Critical Information Needs (CINs) framework defined by Friedland et al., (2012). While detailed and rigorous, this approach, as implemented by many, including Napoli et al. (2017), is also costly and time-consuming and captures only a snapshot of a defined geographic area.

A new approach is needed that integrates the strengths of these prior methods to balance national and regional scope with improved local detail and longitudinal coverage. A semi-automated solution is being developed at the University of Missouri's School of Journalism and Institute for Data Science and Informatics. The project has developed a web crawler to collect news coverage from all accessible sources for a systematic and computer-assisted analysis of CINs and relevant metadata. The software is still in development but, over five months, has collected more than 100,000 articles from 170 Missouri news sources, providing some insights into the solution's viability.

2. The Need for a New Approach

Search and social media platforms account for more than 62% of total digital advertising spend in the US (Grieco, 2020; Myllylahti, 2020), reducing newsroom's financial support and greatly contributing to the "evisceration of journalism" (Pickard, 2020, p. 714). In the last two decades, agenda-setting power has devolved rapidly from traditional media outlets (newspapers, television, and radio) to social media algorithms and consumers. As a result, Facebook and other platforms have been increasing influence in the civic discourse, and Silicon Valley has increasingly extracted and consolidated the revenues generated. In contrast, thousands of geographic areas/counties in the US now lack access to up-to-date local news and information. News deserts are spreading rapidly, challenging the suitability of traditional research methods to keep track. According to a 2018 study (Abernathy, 2018), in four years (between 2014 and 2018), almost 200 of the 3,143 counties in the US have lost their only newspaper.

Hess (2015, p. 486) argued that "local newspapers can…be seen to play a deliberate and active role in generating a sense of community." Studies find that losing a newspaper adversely affects residents' everyday lives and sense of community (Mathews, 2022). Lowrey et al. (2008) provide an ideal starting point for a broader discussion of the relationship between newspapers and community, as the researchers conducted a comprehensive review of mass communication scholarship to analyze "community" and "community journalism," "social glue for the community" and that they "create a shared understanding of what it means to be a member of a community" (Lowrey et al., 2008, p. 284). In recent years, scores of weekly and daily newspapers have vanished from the American news landscape, and thousands of others have become shells, or "ghosts," of their former selves (Abernathy, 2018, p. 24). Though a newspaper may continue publishing under the same name, changes in ownership, staffing, and ambition result in reduced importance and impact in its community (Abernathy, 2018).

2.1. What is Our Method?

The current project explores the use of machine learning (ML) and natural language processing (NLP) to automate the analysis of digital news content along journalism-defined standards. The goal is to collect articles from a comprehensive collection of local news and information providers and use ML models to estimate the extent to which the published material engages in multiple journalistic dimensions, such as spatial and topic coverage, originality, quality, and value to the community. Given the need to gather and analyze journalism from newsrooms nationwide, computational methods are a needed and appropriate solution.

By automating the collection of news stories and using ML to perform an initial analysis of coverage, relevant variables can be tracked over time at any geographic granularity: city, county, state, region, or country—providing an ongoing description of every local information ecosystem. While a promising approach, there are barriers:

1, The complexity of language implies that automated content analysis methods cannot entirely supplant a careful and close reading and manual coding of texts.
2. There are significant challenges in collecting and maintaining the canonical web domains needed to direct the path of the web crawler.
3. The collection of news content is challenging due to paywalls, site design, and inconsistent use of metadata tags.
4. It is not yet clear that the sophistication needed for ML models to accurately identify people, coverage topics, locations, and other metadata is achievable at a threshold needed for rigorous academic study.

However, we argue that the ongoing dynamic changes in the media landscape make it necessary to apply an approach that can incorporate automated data collection, initial ML analysis of big data sets, and human validation and scrutiny of the output. Once collected, this data will allow for timelier and more local insights and provide a new foundation for researchers across disciplines who consider the influence of news ecosystems in the analysis of political, economic, and health outcomes in local communities.

3. The Study of News Deserts

News deserts, defined as areas lacking local news coverage, present significant challenges for the communities affected and researchers studying the phenomenon (Abernathy, 2018). Various methodologies have been utilized, including mapping, surveys, interviews, data analysis, and case studies. Researchers have used Geographic Information System (GIS) mapping tools to visually represent the distribution of news outlets across a region and identify underserved areas (Abernathy, 2018, 2020, 2022; Ferrier et al., 2016; Lee & Butler, 2019; Napoli et al., 2017, 2018, 2019; Stonbely et al., 2019; Stonebraker & Green-Barber, 2021). This helps to understand the geographical patterns of news desert formation and identify areas at risk of becoming news deserts in the future (Napoli et al., 2017, 2018, 2019). Surveys have been conducted to gather data on the news consumption habits of residents, detailing the availability of and access to news in different communities (Shaker, 2014). In-depth interviews with journalists, community leaders, and residents have been conducted to gather qualitative data on the state of local news coverage (Ferrucci & Alaimo, 2020; Stonebraker & Green-Barber, 2021). This provides valuable insights into the experiences and perspectives of people directly impacted by news deserts. Content analysis provides meaningful and scalable measures for the comparative analysis of local journalism across multiple communities or within communities over time (Damanhoury et al., 2022; Neff et al., 2022; Stonebraker & Green-Barber, 2021). Data analysis of circulation and readership of local news outlets, as well as demographics and economic characteristics of communities, has helped to identify patterns and trends in news desert formation and understand the social, economic, and political factors contributing to the lack of local news coverage (Abernathy, 2018, 2022). Finally, case studies of specific news deserts have offered a more nuanced understanding of the factors contributing to the lack of local news coverage in those areas (Ferrucci & Alaimo, 2020).

Each of these methodologies helps researchers better to understand the causes and consequences of news deserts and inform efforts to address this critical issue. Existing research on local journalism and news deserts emphasizes a variety of problems and establishes different approaches. For instance, Napoli et al. (2017, 2018, 2019) and Royal and Napoli (2021) focused on CINs, while Ferrier et al. (2016) examined information ecosystems using GIS data. Damanhoury et al. (2022) examined original, local reporting and coverage of CINs as well as the type of framing in over 600 online stories appearing on the home pages of the site. Abernathy (2018, 2020, 2022) and Stonbely et al. (2019) used GIS, whereas Stonebraker and Green-Barber (2021) used scale-based statistical analysis. Additionally, Ferrucci and Alaimo (2020) used case studies, in-depth interviews, and participant observation to emphasize the influence of community stakeholders on news construction, CINs, and the boundary between traditional and other information sources in a healthy news ecosystem.

Scholars have applied different techniques to study local audiences, their information needs, the closure of news organizations, the rise of ghost newspapers, and the spread of news deserts. Each approach served the research question the scholars wanted to explore or explain but with distinct analytical strengths and limitations.

3.1. Community Information Needs

A convenient starting point from which to consider in news desert research is the Knight Commission on Information Needs 2009 report concerning local information needs (Knight Commission on Information Needs, 2009) and the Federal Communications Commission's 2011 *Information Needs of Communities* (Waldman, 2011).

Also in 2009, the Berkman Klein Center for Internet & Society at Harvard University launched its Media Cloud research database enabling academic researchers, journalism critics, and interested citizens to examine media coverage (Berkman Klein Center, 2009). The database collects articles from selected news sources in the US and internationally but does not provide comprehensive coverage of local markets.

Ferrier et al. (2016) research focused on building capacity for more and better news and information at the local level and developing new ways to connect with the community and catalyze civic responsibility into local solutions. To conduct her initial study, Ferrier et al. (2016) used information from social media, digital ethnography, and narrative mapping techniques that examine geo-specific communities and monitor digital and physical communications. One significance of Ferrier et al. (2016) study was the use of GIS down to the ZIP code level to support the community design of localized solutions. This method allowed examination of the effects of media deserts from three layers: Content (news/information), Code (algorithms, policy, and law), and Conduit (platforms, internet access, and mobile delivery) to model local communication ecosystems. Narrative mapping helps examine the role of digital technologies in creating and sustaining digital identity, social networks, and community engagement.

Napoli et al. (2017) have pursued a series of projects focused on CINs and the coverage and quality of news. Napoli et al. (2017) have led the development of a multi-level methodological framework for assessing local journalism and the extent to which it addresses communities' CINs. The first study examined the journalistic infrastructure in three New Jersey communities.

Napoli et al. (2017, p. 17) described the use of CIN as taking:

> Into account the quantity of journalistic sources located within a community (infrastructure); the quantity of news stories/social media posts produced by these sources, along with the degree of concentration in story/social media post-production (output); and, finally, the extent to which these stories/social media posts meet basic "quality" indicators, such as originality, local orientation, and addressing recognized CINs (performance).

Napoli et al. (2017) used focus group discussions to explore how local news audiences meet their CINs and their attitudes and beliefs about their local news environments. In a following study, Napoli et al. (2018) focused on the characteristics of individual communities and the robustness of the local journalism available to those communities. The study focused on several key concepts: Are some types of communities suffering more than others? Are there particular characteristics of individual communities related to the state of their local journalism? (Napoli et al., 2018, p. 6).

Further, the study used big data sets: To present a rigorous, replicable methodological approach to assessing the robustness of local journalism as we scale the number of communities examined; and to provide descriptive data on the robustness of local journalism by providing indicators of the extent to which local communities are receiving journalism that is original, local, and that addresses CINs (Napoli et al., 2018, p. 3).

In 2018, through geolocation and visualization of newspaper newsrooms, Abernathy (2018) effectively quantified and made visible the national trend of newspaper closures and catalyzed a national policy discussion around the issue. That work was continued by Abernathy (2022), which collected and mapped the locations of more than 8,000 newspapers and digital sites news for analysis.

Damanhoury et al. (2022) compared news coverage from traditional and non-traditional sources across four counties in Colorado. The traditional sources included online stories on the home page of local newspapers, television channels, and radio stations, and non-traditional sources, including the Facebook pages of school districts, government bodies, cities, and NGOs (Damanhoury et al., 2022, p. 7). This study examined how non-traditional media met the CINs of their communities in addition to traditional sources. The research team first reviewed the literature on local news, highlighting CINs originality, locality, and framing as key indicators for assessing the quality of journalism, and underscored the impact of journalistic infrastructure and demographics on news reporting (Damanhoury et al., 2022, p. 2). News from traditional sources was collected for a week in late July 2020. This yielded a sample of 631 stories and a quantitative content analysis. In line with Napoli et al.'s (2017) approach in New Jersey, the selected Colorado counties varied in population size, minority population, income levels, and urban-rural classification. This study was limited in scalability as it depended on manual quantitative content analysis.

Stonbely et al. (2019) looked at the loss of local news coverage, specifically the connection between the community's health and the health of news coverage. The study synthesized prior news deserts and media ecosystem studies to evaluate the gap between citizens' needs and the media's provision. The article attempted to define the theoretical parameters of an ecosystem mapping method that covers a large area while capturing the lived reality of local news ecosystems (Stonbely et al., 2019, p. 1025). In addition, beyond establishing a method for large-scale mapping of local news ecosystems, Stonbely et al. (2019) focused on identifying those communities without regular local government news coverage. The article synthesized the literature and presented a critical analysis of research methods.

Neff et al. (2022) built a multi-dimensional framework for assessing local media systems to identify potential gaps in news provision, especially among socioeconomically marginalized communities. The study gathered data on income, education, and age of audiences and coverage areas for 38 news outlets in Philadelphia and conducted a content analysis to gauge how these outlets meet CINs related to the Covid-19 pandemic. This research outlined a methodological approach to assess multiple dimensions of Philadelphia's media system: audience socioeconomics, audience size, news staffing levels, media ownership structures, and news platforms (more recent digital-only entrants in the media system vs. older, legacy outlets that generally combine platforms such as print, broadcast, and digital).

4. A Scalable Method

The strengths and limitations of these methodologies represent a continuum from richly described but locally and time-constrained ecosystem audits to nationally significant analyses that lack a rich local context. The approach being developed at the University of Missouri consists of six steps:

1. Collection and maintenance of state-based lists of web domains of local information sources.
2. Development of a web crawler to traverse those news sources and regularly gather article text and metadata.
3. Creation of a database to store the collected text and associated information.
4. Training of ML models to analyze the text and extract relevant entities, including location, topics, people, and institutions.
5. Integration of GIS data layers to support analysis.
6. A web interface to enable the production of reports and visualizations.

The first four steps are currently in a prototype phase, with the web crawler collecting more than 100,000 stories from 170 Missouri news sites in the winter and spring of 2022–2023. But each step of the data collection pipeline includes some already recognized challenges, and both expected and still undiscovered roadblocks.

4.1. Collecting News URLs

There is currently no authoritative nationwide list of news outlets, at least partly due to the scale of the effort required and the definitional issues involved.

To direct our web crawler, in the spring and fall of 2022, the team collected lists of news sites from The Center for Innovation and Sustainability at the University of North Carolina, the Editor & Publisher Yearbook, the MediaCloud project at Harvard, and a variety of other sources that included news membership organizations including LION Publishers, the Institute of Nonprofit News, as well as state press and broadcast associations. This data collection aimed to identify any website that might be fairly described as a "local information provider.''

Our initial collection of news sites focused on five states: Missouri, Illinois, North Carolina, New Jersey, and New York. The lists include the outlet name, web URLs, geographic coverage area, street address, print circulation (when applicable), and ownership. The results were manually filtered to eliminate duplicates and compiled into a single database.

Though some of the lists of news sites were of recent vintage, errors and omissions were quickly apparent during the first validation process, which involved manually visiting every collected domain within Missouri.

Most critically, some sites had disappeared or stopped updating. These were not added for data collection.

Another challenge: Some newspapers, especially in smaller, rural communities, do not regularly publish their journalism online. Many offer placeholder web pages or link only to a replica edition or a Facebook page. The publication of a replica/PDF excludes the possibility of automated analysis with our method, though we are evaluating the ability to ingest and process PDFs and scan social media pages. These sites were retained in our lists but not added for data collection.

Mergers, acquisitions, and sales by newsgroups (including the Gannett and Gatehouse combination) are frequent but not always exhaustively reported, complicating the work of creating a canonical list of sites for the web crawler and the task of collecting all relevant articles from each site (also suggested by Lindgren et al., 2020). These mergers often lead to a functional combination of websites where a domain name belonging to a newly purchased newsroom now redirects to a section front on the now-parent company's larger news site.

A similar and common complication was the discovery of new sources of local news that had launched recently or had been overlooked on prior lists. When discovered, these sites were added for data collection, but this effort requires almost constant vigilance and ongoing reassessment.

4.2. The Web Crawler

The process to identify and collect news articles utilizes several Python tools: StorySniffer to determine which URLs on each site are likely to be news articles, Newspaper3k to scrape text and certain metadata fields, and BeautifulSoup to scrape fields not easily obtained with the Newspaper3k package.

The first test of the crawler in the fall of 2022 targeted 264 news sites in Missouri. Over 65 sites were immediately identified as "non-viable" for article collection due to one or more of the causes noted previously.

Another 27 sites presented paywall-related barriers, either blocking the web crawler, providing only a headline and summary text, or limiting access to a handful of articles. Additionally, 21 sites presented technical complications to the collection of story text. These issues appear idiosyncratic and may require significant development effort to optimize. However, 170 of the 264 sites were deemed accessible (some of which posed minor paywall issues), and in December 2022 and January 2023, the crawler collected 33,380 news articles statewide. Approximately 25% originated from television news sites, 13% from radio newsrooms, and 62% from newspaper or digital-only sites.

The paywall-related exclusions can be individually remedied. In the next research phase, still focused on Missouri, we plan to subscribe to or negotiate access to these news sources. However, this approach will be increasingly difficult as the project expands to multiple states.

The strictly technical barriers to web crawler access are solvable at scale but will require dozens or hundreds of customized solutions.

The challenges posed by the discovery and maintenance of a canonical list of news sites are not unique to our method, and the automated monitoring approach we are developing is best equipped to identify "ghost" newspaper sites and to easily add and collect data from newly founded or discovered sources at marginal cost.

It is too early to assess the impact of these automated methods having less than comprehensive coverage of any given local market, region, or state. One hundred percent collection—every story from every news provider—is impractical at any scale for the reasons stated here. The evaluation of "how many is enough" will be a factor of the volume of total outlets vs. those accessible, total stories in a market vs. those collected, and many intangible variables, including geographic location, unique coverage, specific communities served, and the expressed value of those stories by community members. Additional quantitative analysis and local qualitative research will be needed to understand and balance these concerns.

4.3. Development of Machine Learning Models

Once collected, article text is analyzed with several software packages: The Natural Language Toolkit to perform initial NLP; Gensim to identify articles that are exact or close matches of one another; and spaCy for named entity extraction of geographic locations.

In addition to spaCy, we evaluated several statistical learning NLP tools that have been used in other journalism-content analysis projects, including the Stanford Named Entity Recognizer, Flair, DeepPavlov, General Architecture for Text Engineering, and Polyglot. Three core criteria were used to evaluate the tools: whether they were open source, whether they processed data quickly, and whether they could be modified easily to extract entities that are commonly unique to news articles. In a comparison of these tools applied to news texts, Vychegzhanin and Kotelnikov (2019) found that the open-source spaCy, Polyglot, and General Architecture for Text Engineering were the fastest, compared to Flair, the Stanford NER, and DeepPavlov. The Missouri team selected spaCy because of the ease of adding custom rules and retraining the model.

In the fall of 2023, we expect to manually code a sample of the 100,000 articles already collected to classify the seven CINs, plus sports (a CIN + 1 formulation used by a team of researchers at Rutgers and Montclair State University; M. Weber, personal communication, 27 July 2022), and to train the ML model using that corpus. The addition of sports to the core CIN reflects the importance of that genre to local news coverage. It is a discrete section in the newspaper, online, and broadcast outlets, and one easily identifiable using ML models. Expansion of additional topics within the CIN framework is anticipated as the broad definition of the current categories complicates automated analysis. For example, the civic information category includes many potential genres and might be subdivided into local politics, community events, and community leaders. Similarly, the emergencies and risks category encompasses crime and natural disasters. But, as a starting point, the team utilized: emergencies and risks, health and welfare, education, transportation; economy, the environment, civic information, and sports

5. Pilot Project

An initial ML analysis was performed in the spring of 2023 as part of a MA thesis at the Institute for Data Science and Informatics, University of Missouri. The project considered the headline, publication date, author, article URL, header image URL, article text, hostname, article tags, and other metadata (Zdun, 2023).

That data was used to develop an automated process to determine the geographic distribution and category of local news coverage in a given market.

For this pilot, three designated market area (DMA) regions in Missouri were extracted from the initial statewide collection. The markets were chosen for geographic and demographic diversity: Joplin, the smallest, in the southwest corner of the state; Springfield, a mid-sized market in the center of the state; and St. Louis, a large market in the eastern portion of the state. The sample included 3,564 unique articles from 46 newspapers, seven television stations, and one magazine.

Named entities were extracted from the articles, geocoded, and plotted on a map. The bigram dictionary and zero-shot learning methods were also used to classify the articles. The baseline model achieved precision scores generally in the 60–70% range and recall scores in the 80–90% range for the subset of articles tested. In this context, precision measures how many articles were correctly classified (of all articles classified as belonging to the category "crime," what share were actually about crime). Recall measures how many articles were found in a given category out of all true articles in that category (of all articles in the sample about crime, what share did the model correctly identify).

The zero-shot learning model performed better (see Table 1), achieving precision scores largely in the 80–90% range and recall scores in the 90% range for the subset of articles tested. The category with the lowest recall score was politics and civic life, indicating that, among all true politics and civic life articles in the subset of articles checked, the model identified 71% correctly. The relatively low recall score was likely in part because politics and civic life is a broad and nebulous category. Weather articles also scored relatively low in precision and recall, likely in part because weather articles took many forms, and so were not easily classified by the model.

Sinha et al. (2022) showed in their comparative analysis of ML techniques for text classification that various supervised learning models produced precision scores between 69% and 91% and recall scores between 73% and 91%. Among deep-learning models, Minaee et al. (2021) showed that mean average precision scores largely rest in the 65% to 80% range, with some reaching 90%. As the zero-shot learning model yielded higher precision scores across all categories than the bigram dictionary approach—and ones well within and, in some cases, exceeding the best-performing scores in other studies—it was used to classify all articles in the sample.

For this study, the zero-shot precision scores in the 80–90% range indicate that the vast majority of articles that were labeled as a certain category did actually belong to that category. The even higher recall scores seen in the road and car crashes, sports, crime, health, community events, and features categories indicate that very few articles that actually belonged to these categories were overlooked or missed in the classification process.

In the collected sample of 3,564 articles, 1,439 were classified as Community Events and Features, 613 as Crime, 551 as Sports, 451 as Politics and Civic Life, 184 as Roads and Car Crashes, 164 as Health, 95 as Fire and Natural Disasters, and 55 as Weather. A small number were unclassified and labeled as Miscellaneous. These categories align with, but do not exactly descriptively match the original CIN framework and were chosen to accurately reflect the volume and specificity of topics observed in the data. As the ML model improves in the next research phase, more work is needed to normalize and then expand the categories within a CIN taxonomy.

The share of the nine topic categories was relatively consistent between the three markets, with no more than a three-percentage point variation—except for crime coverage in St Louis and sports in Joplin, which were significantly higher than the average.

Figures 1, 2, and 3 show the distribution and type of coverage produced by newspapers and television stations for selected CIN categories in the three Missouri regions.

When combined with other data about the underlying region, this distribution and coverage analysis could reveal interesting insights about the characteristics of regions where CIN gaps exist or news is not produced at all. Those overlays were not applied in the pilot project, but beyond basic demographics, researchers are collecting data that include literacy rates, electoral participation, governmental spending, local retail sales, philanthropic investments, and residents' commute times. This data will be used to evaluate both the audience demand for local news and the capacity of the community to economically self-sustain the provision of local information.

5.1. Geographic Distribution of Coverage

Our consideration of a community-focused definition of "media markets" relies on understanding the place-based location(s) of each news article produced in a region. This is to locate the coverage in a local context

Table 1. Precision and recall measurements using the zero-shot model analysis.

	Category	Precision	Recall	F_1
Zero-Shot	Politics and Civic Life	0.9667	0.7073	0.8169
	Roads and Car Crashes	1	0.9474	0.9730
	Sports	0.9583	0.9787	0.9684
	Crime	0.8667	0.9512	0.9070
	Health	0.9091	1	0.9524
	Community Events and Features	0.8992	0.9469	0.9224
	Weather	0.75	0.75	0.75
	Fire and Natural Disasters	0.8571	0.8571	0.8571

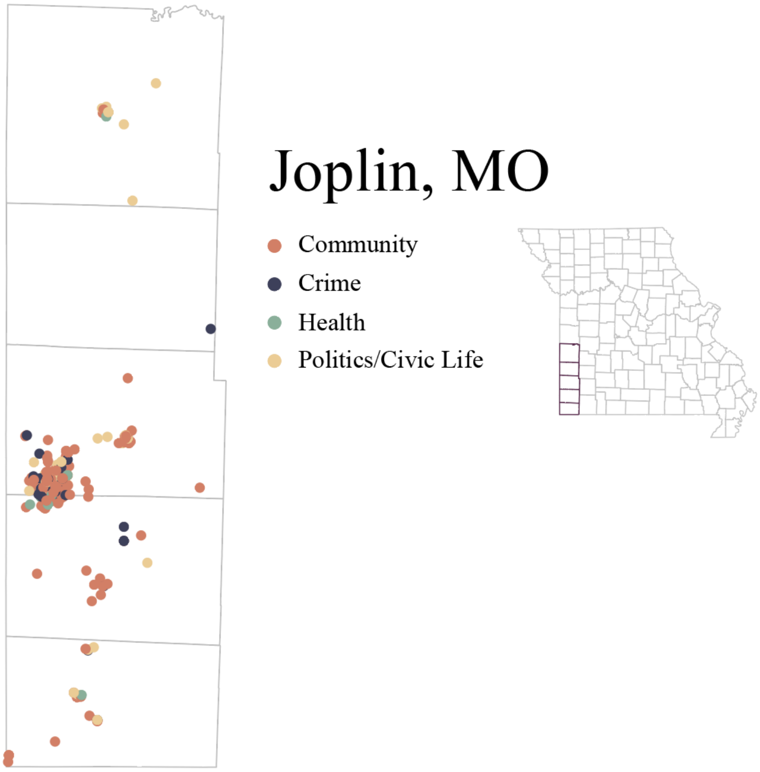

Figure 1. A selection of the CIN categories covered in the five Missouri counties within the Joplin-Pittsburg DMA studied in December 2022.

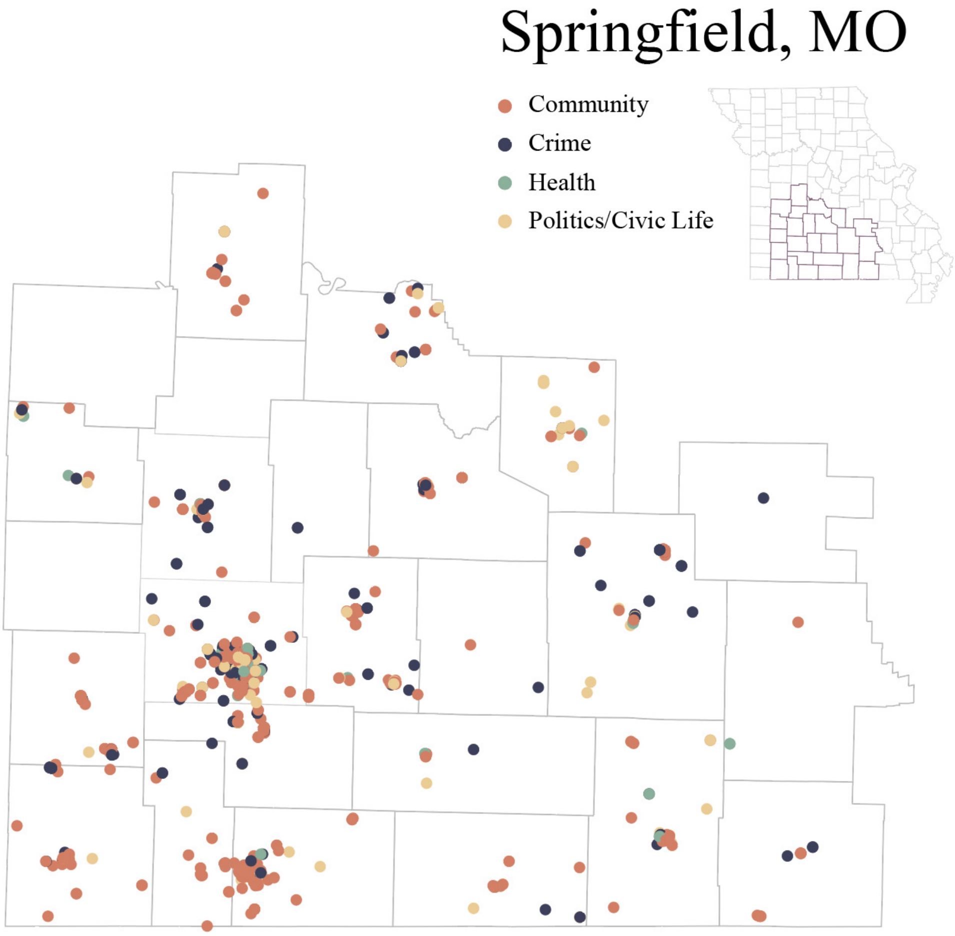

Figure 2. A selection of the CIN categories covered in the 25 Missouri counties within the Springfield DMA in December 2022.

and to filter out locally published stories that refer exclusively to non-local events.

The geographic distribution of coverage in our sample was determined by extracting certain named entities from each article using the open-source spaCy package and its en_core_web_md model, a medium-sized pre-trained model used for tokenization, part of speech tagging, named entity recognition, and other text analysis tasks. For each article, we collected three types of named entities, if they existed: geopolitical entities (such as cities and towns), facilities (such as bridges and airports), and the names of organizations.

The developer also wrote several custom rules using spaCy's rule-based matching feature to collect certain named entities commonly used in news articles but not typically recognized by the pre-trained model. This included block patterns ("4200 block of Maple Avenue"), street patterns ("Gloria Street near Jefferson Avenue"), and other patterns that were unique to news articles.

All extracted named entities from the news stories were then geocoded using the Google Maps API, and the relevant latitude and longitude pairs were plotted on a map to visualize areas receiving news coverage.

5.2. Type of Coverage

The geographic distribution of news can only be evaluated with an understanding of the type of coverage provided in each instance. Under the CIN framework, a healthy information ecosystem requires a balanced mix of news topics. Our research evaluated two data science methods for categorizing news stories: a bigram dictionary and a zero-shot learning method.

More than 900 articles were hand-labeled using the bigram dictionary approach, classifying each into CIN-aligned categories: sports, civic, crime, roads, health, or other. A Term Frequency Inverse Document Frequency Vectorizer was applied to determine the top bigrams for each category. For example, the top three bigrams in the crime category were Police Department, Police Said, and County Police. The top bigrams in the civic category included City Council, School District, and School Board. These bigrams across the six categories were added to a dictionary to classify future articles.

Zero-shot learning, an increasingly popular application of ML, does not require pre-labeled data (Yin et al., 2019). These models evaluate new text on a set of labels not previously seen by the classifier. The open-source

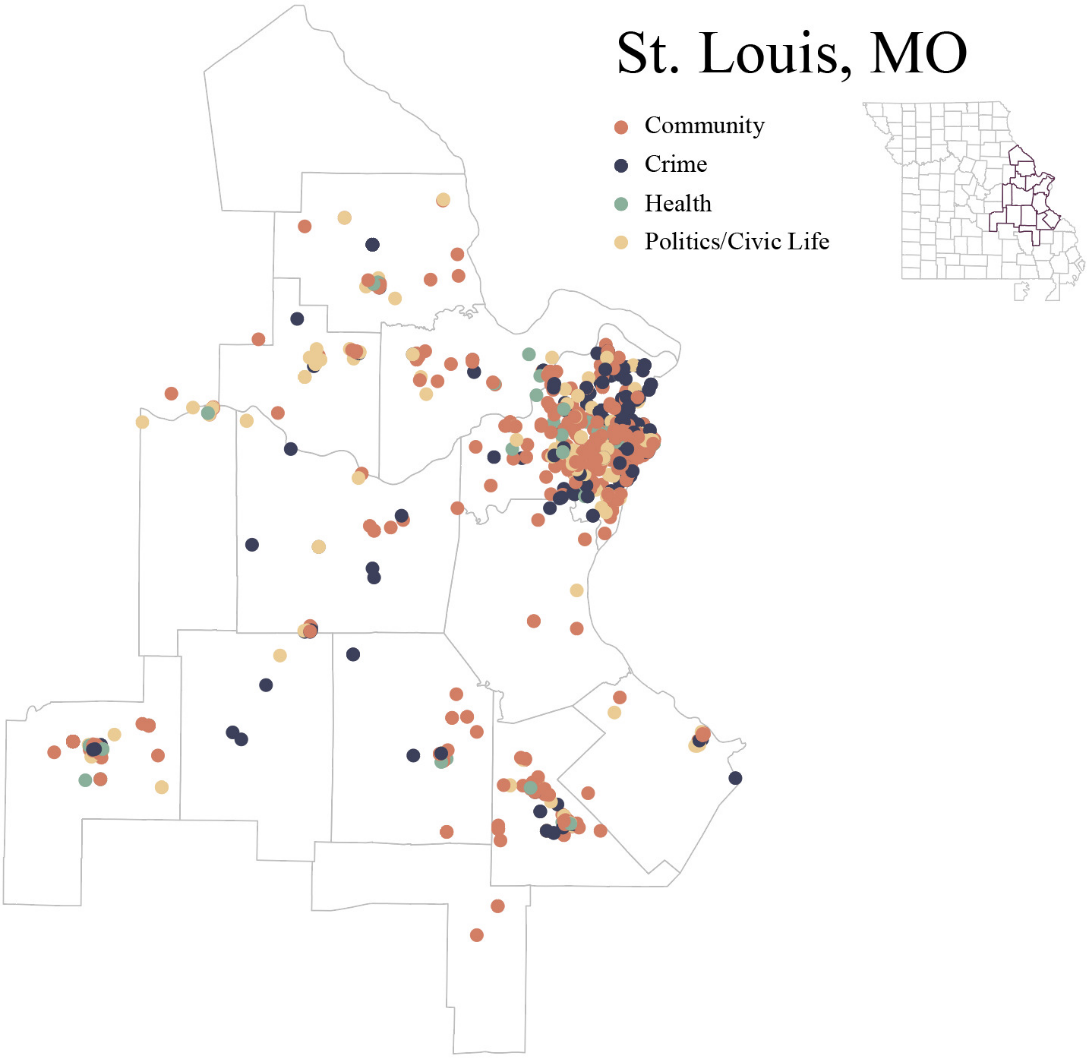

Figure 3. A selection of the CIN categories covered within the 15 Missouri counties of the St.Louis DMA in December 2022.

"bart-lage-mnli-yahoo-answers" model, a transformer encoder-decoder model, was used.

As with other natural language inference models, the zero-shot model compares a hypothesis with a premise. The hypothesis statement used in this case included: "This text is about" as well as nine categories that mirrored the CINs framework: Politics and Civic Life, Community Events and Features, Sports, Roads and Car Crashes, Health, Crime, Fire and Natural Disasters, Weather, and Miscellaneous. To classify the articles, the model determined whether there was an entailment, contradiction, or neither between the hypothesis statement and the premise, which, in this case, was the article text. For example, a short article in the *Bolivar Herald-Free Press* in southeastern Missouri: "Choirs and instrumentalists took the stage at Southwest Baptist University's Pike Auditorium during their 42nd annual Festival of Christmas on Sunday, Dec. 4." The zero-shot model took this article text as the premise, applied the hypothesis statement, tested all of the categories, and ultimately determined that there was an entailment between the article text and the "community events and features" category.

6. Discussion

Despite the robustness, relevance, and value of previous research, the data, tools, and methods used to explore news ecosystems are frequently constrained by geographic boundaries or a lack of rich local detail. This current research introduces automated approaches that can address these limits.

Once fully developed, our web crawler and ML tools will contribute to the study of CINs and news deserts by collecting locally-created news coverage and generating a geo-located analysis of each article text to allow selected CIN categories to be evaluated over time at any level: city, county, state, region, or country—providing an almost hourly snapshot of the health of any local information ecosystem.

News deserts are not a static or binary condition. There is information provision in communities that have no local news outlets, just as towns with a local newsroom may remain critically underserved (Usher, 2023). Moreover, a news desert does not arise overnight or without civic influences and regional or national economic trends (Hess & McAdam, in press). From a distance, news-poor communities may share descriptive characteristics, but the combination of particular causes and solutions is unique to each market.

A quantification of these characteristics: counting newsrooms or local journalists or performing audits of CINs tells the "what." However, that story is incomplete without the qualitative work in a local community to understand the "how" and "why" (Mathews & Ali, 2022). This pairing of quantitative and qualitative methods is expensive, imposing a selection bias on the communities analyzed and effectively limiting the volume of such reports.

This challenge of historical context, data collection, measurement, and analysis, hinders policymakers, funders, and practitioners from effectively intervening to stem a news drought or reverse a desert. Attempts have been made, and some have succeeded, but decisions to invest are driven more by vibes than deeply researched data paired with identified community needs, thereby replicating the errors of inequity that distorted past efforts (Usher, 2021).

Adapting local news ecosystem research methods to enable longitudinal and geographic scales is not without its challenges. A proper understanding of the information health of a single community demands thorough local knowledge and analysis. Data collection and evaluation of thousands of individual news ecosystems across the country requires a scale of work only practical with the support of an automated process. ML and NLP techniques are well suited for this task.

Researchers need access to a near-real-time cross-market evaluation of the quality and coverage of locally generated news to inform policy solutions and evaluate the health of communities, as well as the business of news. Journalism funders need similar insights to understand where and when to invest in legacy newsrooms or to take a bet on new start-ups. And start-up founders have the same questions.

By examining the volume and distribution of topics produced by local journalists, a better understanding of the interaction of local stories and their trajectories in these markets can be quantified and projected. Using network-based time and distance models, the core and periphery areas of news markets can be determined. Geographically weighted regression models can be applied to understand the local news ecosystem and its development. In addition, access to reliable data within and between media markets around the country can support research questions across disciplines, including media ownership, electoral turnout, Covid-19 vaccinations, gun violence, community health outcomes, economic growth, equity in public policy, and climate change preparedness.

Both logistical and technical barriers remain. Maintaining an accurate list of local information providers is difficult, given the rate of openings, closures, and the inherently local nature of this information (Lindgren et al., 2020). We have also highlighted the barriers posed by paywalls, web design, and the unique vocabulary and syntax of news writing. But our pilot project's results have shown the approach's potential long-term merit.

Acknowledgments

Thanks to the entire research team on this project at the University of Missouri, including Dr. Timothy Haithcoat, Dr. Baekkwan Park, Randy Picht, Aaron Babel, and Srijita Datta. This project has received partial funding from the Inasmuch Foundation in Oklahoma City.

Conflict of Interests

The authors declare no conflict of interests.

References

Abernathy, P. M. (2018). *The expanding news desert*. Center for Innovation and Sustainability in Local Media, School of Media and Journalism; University of North Carolina at Chapel Hill.

Abernathy, P. M. (2020). *News deserts and ghost newspapers: Will local news survive?* University of North Carolina Press.

Abernathy, P. M. (2022). *The state of local news*. Local News Initiative. https://localnewsinitiative.northwestern.edu/research/state-of-local-news/report

Ali, C., Radcliffe, D., Schmidt, T. R., & Donald, R. (2020). Searching for Sheboygans: On the future of small market newspapers. *Journalism*, *21*(4), 453–471.

Berkman Klein Center. (2009). *Introducing media cloud*.

Damanhoury, K. E., Coppini, D., Johnson, B., & Rodriguez, G. (2022). Local news in Colorado: Comparing journalism quality across four counties. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2022.2083003

Ferrier, M., Sinha, G., & Outrich, M. (2016). Media deserts: Monitoring the changing media ecosystem. In M. Lloyd & L. Friedland (Eds.), *The communication crisis in America, and how to fix it* (pp. 215–232). Palgrave Macmillan. https://doi.org/10.1057/978-1-349-94925-0_14

Ferrucci, P. (2019). *Making nonprofit news: Market models, influence and journalistic practice*. Routledge.

Ferrucci, P., & Alaimo, K. I. (2020). Escaping the news desert: Nonprofit news and open-system journalism organizations. *Journalism*, *21*(4), 489–506.

Friedland, L., Napoli, P., Ognyanova, K., Weil, C., & Wilson, E. J., III. (2012). *Review of the literature regarding critical information needs of the American public*. Unpublished manuscript. https://transition.fcc.gov/bureaus/ocbo/Final_Literature_Review.pdf

Gabbatt, A. (2020, April 9). US newspapers face "extinction-level" crisis as Covid-19 hits hard. *The*

Guardian. https://www.theguardian.com/media/2020/apr/09/coronavirus-us-newspapers-impact

Grieco, E. (2020, February 14). Fast facts about the newspaper industry's financial struggles as McClatchy files for bankruptcy. *Pew Research Center*. https://www.pewresearch.org/short-reads/2020/02/14/fast-facts-about-the-newspaper-industrys-financial-struggles

Griffin, R. (2018, September 5). Local news is dying, and it's taking small town America with it. *Bloomberg*. https://www.bloomberg.com/news/articles/2018-09-05/local-news-is-dying-and-it-s-taking-small-town-america-with-it

Hayes, D., & Lawless, J. L. (2018). The decline of local news and its effects: New evidence from longitudinal data. *The Journal of Politics*, *80*(1), 332–336. https://doi.org/10.1086/694105

Hess, K. (2015). Making connections: "Mediated" social capital and the small-town press. *Journalism Studies*, *16*(4), 482–496.

Hess, K., & McAdam, A. (in press). Degradation and "desertification" of digital news ecosystems. In S. Eldridge, J. Swart, S. Banjac, & D. Cheruiyot (Eds.), *The Routledge companion to digital journalism studies*. Routledge.

Hutchins, C. (2022). *Introducing the Colorado news mapping project*. Colorado News Collaborative. https://colabnews.co/projects/colorado-news-mapping-project-method

Jordon, S. (2018, April 13). Shareholders won't be tossing newspapers at this year's Berkshire Hathaway meeting. *Omaha World-Herald*. https://www.omaha.com/money/shareholderswon-t-be-tossing-newspapers-at-this-year-s/article_9a9b2a84-9dfd-5638-bc39-3d27b7b4ffe8.html

Knight Commission on Information Needs. (2009). *Informing communities: Sustaining democracy in the digital age*. https://knightfoundation.org/reports/informing-communities-sustaining-democracy-digital

Lee, M., & Butler, B. S. (2019). How are information deserts created? A theory of local information landscapes. *Journal of the Association for Information Science and Technology*, *70*(2), 101–116.

Lindgren, A., Corbett, J., & Hodson, J. (2020). Mapping change in Canada's local news landscape: An investigation of research impact on public policy. *Digital Journalism*, *8*(6), 758–779.

Lowrey, W., Brozana, A., & Mackay, J. B. (2008). Toward a measure of community journalism. *Mass Communication and Society*, *11*(3), 275–299.

Mathews, N. (2022). Life in a news desert: The perceived impact of a newspaper closure on community members. *Journalism*, *23*(6), 1250–1265.

Mathews, N., & Ali, C. (2022). Desert work: Life and labor in a news and broadband desert. *Mass Communication and Society*. Advance online publication. https://doi.org/10.1080/15205436.2022.2093749

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning—Based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.

Myllylahti, M. (2020). Paying attention to attention: A conceptual framework for studying news reader revenue models related to platforms. *Digital Journalism*, *8*(5), 567–575.

Napoli, P. M., Stonbely, S., McCollough, K., & Renninger, B. (2017). Local journalism and the information needs of local communities: Toward a scalable assessment approach. *Journalism Practice*, *11*(4), 373–395. https://doi.org/10.1080/17512786.2016.1146625

Napoli, P. M., Stonbely, S., McCollough, K., & Renninger, B. (2019). Local journalism and the information needs of local communities: Toward a scalable assessment approach. *Journalism Practice*, *13*(8), 1024–1028.

Napoli, P. M., Weber, M., McCollough, K., & Wang, Q. (2018). *Assessing local journalism: News deserts, journalism divides, and the determinants of the robustness of local news*. Dewitt Wallace Center for Media and Democracy. https://dewitt.sanford.duke.edu/assessing-news-media-infrastructure-report-released

Neff, T., Popiel, P., & Pickard, V. (2022). Philadelphia's news media system: Which audiences are underserved? *Journal of Communication*, *72*(4), 476–487. https://doi.org/10.1093/joc/jqac018

Pickard, V. (2020). Restructuring democratic infrastructures: A policy approach to the journalism crisis. *Digital Journalism*, *8*(6), 704–719.

Royal, A., & Napoli, P. M. (2021). *Local journalism's possible future: Metric media and its approach to community information needs*. Researchgate. https://rb.gy/cv3p2

Shaker, L. (2014). Dead newspapers and citizens' civic engagement. *Political Communication*, *31*(1), 131–148.

Sinha, A., Naskar, M. N. B., Pandey, M., & Rautaray, S. S. (2022). Text classification using machine learning techniques: Comparative analysis. In B. Sahoo & S. P. Mohanty (Eds.), *2022 OITS International Conference on Information Technology (OCIT)* (pp. 102–107). IEEE.

Stonbely, S., Konieczna, M., & Holcomb, J. (2019). *Mapping local news ecosystems, phase 1: Meta-review of the literature, a typology of ecosystem studies, and a proposed method for the ideal ecosystem study (Mapping Local News Ecosystems)*. Center for Cooperative Media.

Stonebraker, H., & Green-Barber, L. (2021). *Healthy local news & information ecosystems: A diagnostic framework*. Impact Architects. http://files.theimpactarchitects.com/ecosystems/full_report.pdf

Usher, N. (2019). Putting "place" in the center of journalism research: A way forward to understand challenges to trust and knowledge in news. *Journalism &*

Communication Monographs, *21*(2), 84–146. https://doi.org/10.1177/1522637919848362

Usher, N. (2021). *News for the rich, white, and blue: How place and power distort American journalism*. Columbia University Press.

Usher, N. (2023). The real problems with the problem of news deserts: Toward rooting place, precision, and positionality in scholarship on local news and democracy. *Political Communication*, *40*(2), 238–253.

Vychegzhanin, S., & Kotelnikov, E. (2019). Comparison of named entity recognition tools applied to news articles. In *2019 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 72–77). IEEE.

Waldman, S. (2011). *Information needs of communities: The changing media landscape in a broadband age*. Diane Publishing.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In S. Brewster & G. Fitzpatrick (Eds.), *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–12). ACM.

Zdun, M. (2023). *Assessing the quality and distribution of news in local television and newspaper markets using data science methods* [Unpublished Master's thesis]. University of Missouri.

About the Authors

Asma Khanom is a PhD student at the Missouri School of Journalism, University of Missouri. She is interested in changes and risks in journalism, local journalism, community information need, organizational shifts, media management, and ethical decision-making. Asma has over a decade of television journalism experience in Bangladesh, Germany, and the US. Asma aims to impact existing journalism practices through her research and make a bridge between the industry and academia.

Damon Kiesow is the Knight chair in journalism innovation at the Missouri School of Journalism and a digital media pioneer who specializes in aligning community information needs and business strategy in support of sustainable local journalism. He is a 30-year veteran of local and national media outlets, as a reporter, photographer, editor, and product manager.

Matt Zdun received his MA in data science and data journalism from the University of Missouri. His primary research interest was using data science skills to analyze news content. Previously, he worked as a political journalist in Texas at the Austin American-Statesman and The Texas Tribune. He has also worked in broadcast journalism, at CNBC and a CBS station in Central Texas. He received dual BA degrees from Northwestern University in journalism and economics.

Chi-Ren Shyu is the Shumaker Endowed professor and director of the University of Missouri Institute for Data Science and Informatics. As a National Science Foundation career awardee and a fellow of the American Medical Informatics Association. He has contributed to the fields of biomedical informatics and computational sciences, with over 120 journal publications. Shyu's collaborations with Missouri's Donald W. Reynolds Journalism Institute focus on combatting misinformation and promoting community information equity, employing emerging technologies such as the blockchain and geospatial artificial intelligence.