Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

# Media Cloud 2.0: An Updated Open Web News Archive

**Fernando Bermejo**[*1], **Rahul Bhargava**[*2], **Phil Budne**[*1], **Paige Gulley**[*1], **Evan Leon**[*2], **Ryan McGrady**[*3], **Emily Boardman Ndulue**[*1], **Ethan Zuckerman**[*3]

[1]Media Ecosystems Analysis Group
[2]Northeastern University
[3]University of Massachusetts Amherst
{fernando,paige,emily}@mediacloud.org, {r.bhargava,e.leon}@northeastern.edu,
phil@regressive.org, {rmcgrady,ethanz}@umass.edu

**Abstract**

We present a completely re-engineered Media Cloud, a massive searchable open source archive of digital news sources and content from around the globe. Since its previous presentation at ICWSM in 2021, the Media Cloud team has re-engineered the tool's data collection, storage, and retrieval systems, built a new front-end research interface, surpassed 1.8 billion stories, and reprocessed all the content to update the extracted metadata with consistent and modern techniques. In this paper we document the new system's engineering, characterize the datasets to date, and describe user-facing tools. This includes a Directory of online news sources and a searchable Story Index of global news stories. We discuss the utility of the datasets, how they compare to other related work, challenges associated with maintaining open research infrastructure, and research made possible through the datasets and tooling.

**Geographic Collections Dataset —**
https://doi.org/10.5281/zenodo.18189311

## 1 Introduction

News reflects a society's composition, the discourse through which it understands itself, the events that transpire within it, and the spatial and temporal contexts in which those events take place – the proverbial who, what, when, where and why. Global trends have led to a profound decaying of shared truth, a turn away from facts to feeling in policy analysis and decision-making (Kavanagh and Rich 2018). But even as fewer readers get their news from journalism and more from social media (St. Aubin and Liedke 2025), online news reporting offers a powerful proxy and fuel for what is amplified into social conversation platforms. Researchers across disciplines use news corpora and computational methods to generate important knowledge about narrative amplification (Elfes 2025), health messaging (Qian et al. 2025), partisanship (Chen et al. 2022), and a wide range of other issues and topics.

However, across the globe, news ecosystems are increasingly fragmented across platforms (Fletcher and Nielsen 2017), paywalls (Dhillon, Panda, and Hemphill 2025), media partisanship (Borah et al. 2024), and are constantly evolving. This has created barriers to curating contemporary and usable lists of media sources, and collecting, processing, and storing a news corpus for study. Commercial solutions exist, but are typically expensive, suffer from limited geographic scope, and often have insufficient documentation and metadata to support rigorous computational analysis. The Media Cloud news source Directory and online news Story Index have addressed these concerns since 2008 (Roberts et al. 2021). The Media Cloud tools offer access to curated lists of media sources based on geography and topic, and a searchable web-based interface to over 1.8 billion stories in the online news Story Index. Over the last few years, Media Cloud has emerged as a critical dataset for computational social science, with over 167 academic publications citing their use of the datasets, API, and web-based research interface.

Yet the challenges of building and maintaining public data infrastructure for almost 20 years are complex. Media Cloud's datasets are rooted in work first published in 2004 (Zuckerman 2004), predating the contemporary media environment and the growth of social media. The data collection and management system grew from a set of connected Perl scripts, to a PostgreSQL database, to an interconnected web of dozens of Python micro-services. This unplanned scaling and development led to data quality inconsistencies and brittleness, particularly as story volume exceeded 1 billion stories archived.

This paper documents new Media Cloud datasets that bring consistency in both story and news source data, updated and normalized metadata extraction, and scalability to the service as it continues to grow. This included a complete rewrite of the data ingestion pipeline, a tighter focus on core metadata that could be extracted for all stories, re-processing of the entire historical archive to ensure methodological consistency, developing new front-end research interfaces, and re-categorization of the entire Directory for simplification

[*]These authors contributed equally.

and removal of duplicate sources. The legacy system presented at ICWSM in 2021 was officially disabled in favor of the new system in December 2023, with historical reprocessing of all data completed in September 2025. In this paper we describe the motivation for re-engineering the datasets and software, characterize and describe the datasets offered, explore challenges for sustaining dataset archives such as this, and catalog existing academic use.

## 2 About Media Cloud

Media Cloud was started at the Berkman (now Berkman Klein) Center for Internet at Society at Harvard University in the early 2000s with the goal of providing empirical data to researchers interested in studying the networked public sphere. In 2011 most of the development of the platform was transferred to the Center for Civic Media at MIT's Media Lab, while the Berkman Center kept a leading role in research innovation. Starting in 2016, the platform became global in scope (with media collections from every country in the world) and open to all users (with a user-friendly front-end and sign-in process). The non-profit Media Ecosystems Analysis Group (MEAG) was incorporated in 2017 to serve as the project's fiscal sponsor and manage day-to-day operations. In 2021, the current governance structure of Media Cloud took shape: the datasets and infrastructure are now managed as a consortium between the MEAG non-profit, the Initiative for Digital Public Infrastructure at University of Massachusetts Amherst, and the School of Journalism at Northeastern University. In 2024 an advisory board was created to provide strategic advice. Since its founding, Media Cloud has been supported through a variety of public and private grants. MEAG has also contributed since 2017 to the maintenance of the project through the resources it generates.

### 2.1 Re-engineering and Re-indexing the Datasets

As noted above, it became clear in 2021 that the engineering pattern of the legacy system and the increasing volume of stories archived was not a sustainable combination for Media Cloud. One driver was that the original data storage architecture wasn't designed to scale to the volume of content it grew to. For instance, the full extracted text of stories was being stored in both PostgreSQL and Solr in order to take advantage of the complementary strength of each. However, this redundancy doubled the storage growth rate. Additionally, the set of metadata stored for each story grew in parallel. At a lower level, the primary key space architecture wasn't designed to track unique IDs for the number of rows needed. The data pipeline was also not designed to reduce regular maintenance work. At various points during development of the legacy tool it was conventional wisdom to build for cloud-based deployment, yet decisions we made based on that turned out to be poor fits for academic settings where power and bandwidth are free, but funding to pay the monthly cloud costs might ebb and flow. Limitations like these were the results of organic growth, thus necessitating a re-architecting focused on balancing storage and content needs, prioritizing monitoring and alerting, and detecting data processing edge cases.

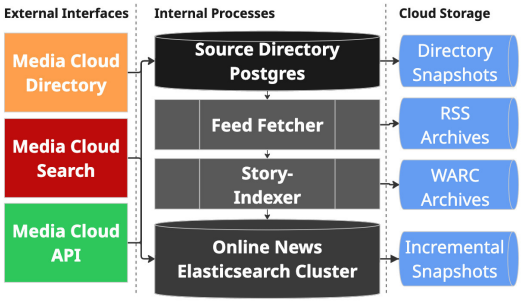


Figure 1: High-level diagram of the key components of the Media Cloud software stack.

We engineered a new ingestion and storage system based on these goals, with production use beginning in early 2023. The reingestion of existing data relied on legacy PostgreSQL databases and backups of RSS and HTML files stored by the legacy system in an S3 bucket (87TB). Based on a review of query logs and a desire to make a contiguous record available, we decided to begin data capture into the new system while in parallel reindexing in reverse chronological order. This allowed recent data to be available more quickly than historical archives. In September 2025 we completed the reingestion process.

### 2.2 New System Model

The Media Cloud technology stack is built on two datasets: the media source Directory and the searchable Story Index of online news. Tying these together is a single web application with two separate data ingestion systems, and incremental off-site backups for each service (Figure 1). This architecture allows for pseudo-real-time capture of online news stories; content shows up in the Story Index for use roughly a day after publication in a source we support.

For sources in the curated Directory, software discovers feed URLs that each uses to syndicate their stories to other sources (supporting both RSS and Google News sitemaps). The Feed Fetcher adaptively polls those feeds to discover candidate story URLs, storing them for later download and processing (and archiving the URLs publicly). The Story Indexer consumes those stored story URLs and queues them for fetching, parsing, and indexing. Data in the Directory and Story Index are made available via web-based tooling and an authenticated API. The content, metadata and database are regularly backed up to offsite storage.

An overarching design principle of the new Media Cloud system model is simplicity; key decisions were made to avoid over-engineering, leverage existing open source solutions, and operate at scale. The new system replaces relational databases with message queues for better parallel processing and abstracts wrapper objects to encapsulate story data. Quarantine queues are used to handle consecutive recoverable failures (invalid redirect, parse error, etc.). These approaches build on years of accrued knowledge about the legacy system. At current growth rates we have capacity for roughly five years of continued ingestion of new content.

**Source Directory** The Media Cloud Source Directory is manually curated database of online news publications. It takes the place of the Source Manager in the legacy system, with a key differentiator being that each source now represents a unique domain (with a few exceptions), as opposed to a source representing the stories from a set of associated RSS feeds. The Directory currently contains over one million entries, though only 3% of those have more than 1,000 associated stories in the Story Index.

The Directory automatically discovers and tracks syndication feed for sources, capturing RSS or Google News Sitemap files offered at the source's domain (via our public `feed_seeker` Python package). The Directory currently tracks almost 150,000 live feed URLs, though only about 50% of those respond to requests with valid feed data.

A key change vis-a-vis the legacy system is that the source record is defined as unique almost exclusively by the root domain. The legacy system allowed multiple sources to reference the same domain, linking them back in the searchable indexes by database ID values. If created properly, with appropriate feeds supplying subsection content, this allowed researchers to investigate one portion of a publication, such as New York Times op-eds or sports content. Over time this design did not scale well, and did not evolve alongside research practices that have grown since the blog era of the internet. The re-engineered Media Cloud now treats a source as a domain, while also allowing for more fine-grained and slower searching via url-based patterns when required.

There are a handful of exceptions to this architecture. Providers such as Wordpress are supported via an optional URL-based search pattern that can be added to a source. This allows advanced users to limit a source to stories at a shared domain ("www.wordpress.com") that only match a specific URL path ("somervillenews.wordpress.com"). Another feature supports sources that have migrated domains. This alternative domain system allows us to associate sources such as the Huffington Post, which was originally published at "huffingtonpost.com", but then migrated to "huffpost.com".

Collections are curated groupings of sources based on clear and documented parameters, such as geographic location of the sources, target demographic of the readership community, or topic of reporting. Collections can also be created based on external research into media groupings, such as media ownership or media partisanship slant. There are varying levels of permissions to allow for shared curation of collections.

The collections system is intentionally flexible, offering a nimble approach to support evolving research uses of the Directory and Story Index. Collections include highly curated internal research products, copies of externally produced sets, one-off speculative lists, experimental leftovers, and more. One key contribution is the curated set of global geographic collections, which can be used for other research and media analysis systems. We collaborate with external partners, and staff a digital librarian as we are able, to maintain and update these collections as time and financial resources allow.

**Data Ingestion Process** Data discovery is initiated from the Directory's list of feed URLs. The Feed Fetcher dynamically retrieves both public RSS feeds and Google News Sitemaps. The code adaptively adjusts the interval between fetches of feeds, aiming to keep the number of new story URLs between 33% and 66% of the total number of URLs in the feed document. This heuristic was iteratively determined by attempting to balance the goals of respecting the host's server capacity and our desire to ingest comprehensive content. Based on maintenance needs, the service also supports fetching on demand and ensures only one live request at a time. Candidate story URLs surfaced from feeds are stored and archived publicly.

The Story Indexer is a flexible queue-based pipeline for consuming URLs from various sources and indexing their content. It's primary task is to consume story URLs from the Feed Fetcher and process them for download, parsing and indexing. Alternate data flows we created allow for re-ingestion of content from the legacy Media Cloud system or large batches of CSV files. No matter where the originating URL comes from, the resulting metadata from the Story Indexer is stored in the Elasticsearch index.

Various back-off and retry strategies attempt to balance our archiving focus with respect for the public content servers we make requests against. For example, fetch failures from any source are re-queued after a one-hour delay (and discarded after excessive retries), and URLs that would exceed per-domain rate limiting are re-queued after a short delay. We endeavor to use a single User-Agent string in all web fetches to identify the project and its academic status.

**Online News Story Index** Stories fetched and processed get stored in the Story Index, our core data store. As of December 2025 this database stores more than 1.8 billion documents in an Elasticsearch cluster of 8 nodes self-hosted on our in-house hardware. Tuning in the hardware and software requirements of this Elasticsearch database was one of the more complex questions we had to answer on the path to our current product; the legacy system was hosted on an Apache Solr cluster of only three rather large nodes. Our choice to spread the compute horizontally across several machines, along with carefully adjusting a great many configuration parameters, resulted in improved performance and reliability against internal metrics, along with a clear plan for ongoing maintenance as the Story Index continues to grow. The standard Elasticsearch APIs allow data search and retrieval.

The Story Index does not permanently keep any story data or metadata in a relational database; all searchable data is kept only in Elasticsearch. The legacy system duplicated this content in a relational database, originally designed to support sentence-level searching. However that feature - which required storage of individual sentence records in PostgresSQL - was removed for scaling needs years prior to discontinuing the legacy system. All data is stored in duplicate in the cluster, ensuring the system can survive the loss of a single node. Elasticsearch snapshots are taken twice a month and stored offsite. All data about news articles, including the extracted text, the original HTML and all extracted metadata are written to compressed Web ARChive

(WARC) files and archived to cloud storage (International Organization for Standardization 2017).

### 2.3 Data Access

To accommodate researchers and more lay users, the Directory and Story Index datasets are accessible via a web interface and API endpoints. These access both the Postgres Directory database and the Elasticsearch Story Index. Full text of articles is not available publicly in either interface due to copyright concerns. Our aim is to provide a system that allows for careful study of news stories while preventing their wholesale copying and reproduction. This has become even more important in the age of large language models and related concerns and lawsuits from news organizations.

Searches against the Directory of media sources and collections are term-based matches against the publication name (extracted from the homepage or manually created), collection name (manually created), and canonical domain name (system-generated).

Searches against the Story Index of online stories are boolean combinations of search terms in a reduced form of the Elasticsearch query syntax. At minimum a query matches the full text content of stories against a search string, a range of publication dates, and a list of media source domains (derived from the sources and/or collections offered in the Directory). It is possible to add additional constraints against any field in our database (such as the article's title, language, etc.). The core search methods can return the full list of matched stories, but can also return sub-samples and aggregations over the key fields to allow for simple time series analysis, top words in article titles, etc. These core data access methods are centralized in an internal software library used in various system components.

**Web Search** The legacy Media Cloud system included two research tools, Explorer and Topic Mapper, which were continuously expanded and built out with features. Unfortunately over time this led to a hard-to-maintain set of brittle tools. This necessitated the development of the new web search platform, providing researchers with accessible interfaces to query the Story Index that accommodate users with varying levels of technical expertise. The web-based interface offers two search modes: a simple search interface for users unfamiliar with boolean query construction, and an advanced search mode for those comfortable authoring more complex boolean queries. Both interfaces allow users to select media from the Directory by choosing any set of individual sources and/or collections. Lastly, users specify a date range for their analysis. Additionally, users can author multiple queries to compare the results.

Search results are presented through multiple complementary views. A time series of content summarizes content over time, displaying both the raw count of matching stories per day and the percentage of total stories published that day matching the query by the selected media outlets. Total attention metrics aggregate these patterns across the entire search period, showing both absolute story counts and the percentage of all stories published during that timeframe by the selected media outlets. A sample list of matching stories provides direct access to individual articles, while other aggregated views surface the top languages, most frequent words, and most active sources within the result set.

All data presented in the web interface is available for download as CSV files, including complete lists of URLs for matching stories and the underlying data powering each visualization.

**API** All eight of the data views provided by the web search platform can also be accessed via the API, which wraps these read-only endpoints to allow more technical users to integrate our data into third party systems. We maintain an open-source `mediacloud` package to allow for easier access in Python, and provide documentation and examples in the form of Jupyter notebooks.

We gate usage behind a token authentication system, requiring users to verify their accounts within our system and acknowledge their understanding of our terms of service via email verification. In addition, each account is offered a quota of searches, reset weekly. This authentication pattern has helped us isolate and alleviate problematic usage, while still leaving the platform as open as possible. Our team regularly provides research users higher levels of quota access based on requests describing the research use and need. Our terms of service forbid (though don't effectively prohibit) signing up for multiple accounts in order to work around quota limits. Experience repeatedly demonstrates that the token-based access, duplicate account prohibition, and quota system are critical tools to prevent abuse of the system.

**Adherence to FAIR principles** Media Cloud is committed to ensuring that the database adheres to the principles of being Findable, Accessible, Interoperable, and Reusable (FAIR) (FORCE11 2020). We promote being Findable through unique and persistent story and source identifiers, well-described metadata fields, and via the core searchable Story Index itself. We promote being Accessible through a document search API and provide a Python client implementation and a simpler web-based search interface available to anyone willing to register with their email. We promote being Interoperable through publishing our source Directory to data archives, offering a public daily file listing all news story URLs we discover, and relying internally on well-defined protocols and formats such as WARC files. We promote being Reusable through offering CSV downloads, API-based access in JSON format, keying content provenance to a public URL, and describing data formats and protocols in online documentation and papers such as this.

## 3 About the Datasets

The Media Cloud project provides an open-source set of technologies that support researchers generating their own corpora of media sources and/or online news content for study. It is worth noting that the comprehensiveness and coverage is biased towards the topics and geographic areas of study of the core team and contributing colleagues; under- or yet-to-be- studied topics and regions may not be well represented within Media Cloud. This section quantitatively summarizes and characterizes the two datasets.

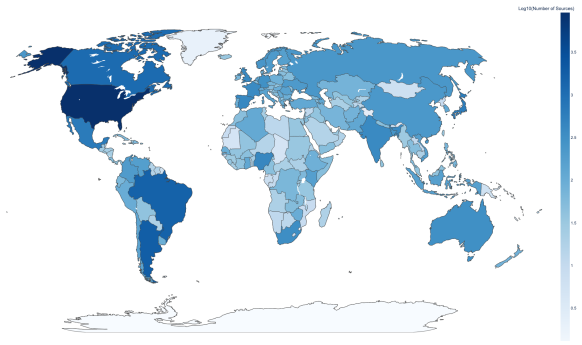

Figure 2: The number of media sources in the directory by country of publication (log scale).

### 3.1 Media Source Directory

The media source Directory holds unique domain names of news media, which we refer to as sources. Our Directory is built to track news-focused websites that regularly publish general interest stories as web pages. This includes global news sites, hyper-local publications, and trade journals, but currently excludes niche blogs, marketing sites, and content syndicators.

**Collections** As of December 2025 the Directory contains 1,931 curated collections of news sources. Collection size varies substantially: the mean number of sources per collection is 78.1, while the median is 6, reflecting a long-tailed distribution in which a small number of large collections coexist with many narrowly scoped ones.

Driven by research interests that often focus on geographical domains, we created a large set of curated country- and province-based media collections. This includes over 1.5 thousand collections made up of over 28 thousand media sources. A static snapshot is available at https://doi.org/10.5281/zenodo.18189311. Figure 2 shows that those sources are distributed globally, with a high concentration in the United States (the home territory of the project). The quality, vetting, and timeliness varies based on which have been used for research recently; a free text "notes" field captures changes and activity on each in order to offer a light historical record of changes and credit to contributors.

**Sources** A media source represents a unique domain publishing content. As of December 2025 there are approximately 26,000 live sources (among the 1,147,442 sources in the system), which collectively publish through more than 105,000 live feeds. In this case "live" refers to a source or feed that provides new content regularly. A significant portion of the previously active sources were blog pages (such as numerous LiveJournal accounts) that we decided to stop collecting data from for efficiency reasons.

A significant portion (95%) are "placeholder" sources that have less than 100 stories (Table 1). These are primarily legacy sources that were part of the spidering-based Topic Mapper legacy tool, captured to support analysis on a topic rather than to track content over time. With limited compute and storage capabilities, we don't automatically try to pull all stories from every source such as these in our system. Four large news aggregators (blogdomago.com, goo.ne.jp, google.com, yahoo.com) provide a significantly out-sized number of stories into the Story Index, but these are not incorporated into the curated research collections unless specifically selected by a user.

Media sources include a number of metadata fields to describe attributes that might support research (Table 2). Adding and maintaining this data is a time-consuming manual process, so completeness is sparse across the live sources. For example, 41% of live sources have a publication country, 27% have a publication state, and 38% have a media type set.

### 3.2 Online News Index

As of December 2025, the Story Index of online news stories contains over 1.8 billion stories, growing at a rate of almost 400,000 stories a day. In order to limit the growth of storage needs over time, we store a minimal representation of each story's metadata as a document in the core Elasticsearch database. We precompute three additional fields of metadata beyond the details of the article itself: the article's title, the publication date, and a best guess as to the language of the article. The text content itself is not available to users for export out of an abundance of caution regarding copyright law, but is kept as the primary search field. See Table 3 for the metadata fields indexed on each story. The software implementation of each is captured in a single `mediacloud-metadata` software package published to PyPI. Many of the underlying methods rely heavily on the trafilatura software library (Barbaresi 2021).

Our indexing has grown over time as new research projects and funding have started and stopped. Figure 3 shows the growth in the total volume of the Story Index over time.

Rapid increases in volume parallel investments in maintenance and expansion of our Directory of media sources and collections. For example, in 2013 the project began indexing content from sources contained in the European News Monitor, and then in 2017 a large batch of sources from the ABYZ News Links project were added. We are uncertain of the cause of the overall decline in story volume since 2022. This could be the result of both the rise in stricter content controls across the web in the era of generative AI, or per-

| Story Count | Number (Percent) of Sources |
|---|---|
| 0–9 | 967,640 (84%) |
| 10–99 | 113,174 (10%) |
| 100–999 | 35,442 (3%) |
| 1,000–9,999 | 17,200 (1%) |
| 10,000–99,999 | 10,574 (1%) |
| 100,000–999,999 | 3,228 (<1%) |
| 1,000,000–9,999,999 | 179 (<1%) |
| 10,000,000–99,999,999 | 4 (<1%) |

Table 1: Distribution of story count by sources, including live and not-live sources.

| Attribute | Description | Notes |
|---|---|---|
| Id | A unique identifier within our system. | Generated automatically. |
| Homepage | The full URL to the source's homepage. | Created by the user; used to run automated feed discovery. |
| Domain | The root domain and suffix of the site. | Treated as a canonical identifier for the media source. Extracted from the user-supplied homepage value. |
| Label | A human-readable name for the publication. | Defaults to the domain, but can be extracted from the homepage if scraped for feeds, or overridden by user input. |
| Publication Country | The country the source is published from. | User-generated as a 3-letter ISO 3166-1 alpha-3 value. |
| Publication State | The state or province the source is published from. | User-generated as a hyphenated ISO 3166-2 value. |
| Media Type | A description of what type of source it is, one of "digital native", "print native", "audio broadcast", "video broadcast", or "other". | User-judged based on manual coding criteria. |
| Notes | Open text list of changes, credits, and related items useful for a user of the source to be aware of. | Highly variable content, but often includes dated credits to individuals or organizations. |
| URL Search String | A prefix-matched wildcard string for URL-based sub-searches designed to support publications that are published at the same root domain. | User-generated on sources treated as "children" of the main source. |
| Stories per Week | Weekly average number of stories published over the last few months. | System-generated via manually-triggered analysis. |
| Primary Language | The language most often detected on related text (2-letter ISO 639-1). | System-generated weekly for new sources that had had any stories indexed in the last 6 months. |

Table 2: Metadata about sources included in the Directory

haps related to the decline of sources generating RSS feeds. This issue warrants more investigation that is out of scope for this dataset paper.

A diversity of global languages are included in the archive, though more than 44% of the stories are in English (Table 4). 17% of the stories are detected to be written in languages not represented in that chart. While the dataset includes content in over 100 languages, it does not actively support searching logographic languages due to technical constraints.

## 4 Related Work

There are three frequently overlapping groups of projects or tools that are similar in scope to Media Cloud: directories of news sources, news databases, and event databases.

### 4.1 Directories of News Sources

Directories of newspapers and other media have existed since the earliest days of the web, often with a straightforward HTML structure that has made them easy to use for research purposes. Two examples that still exist are ABYZ News Links and OnlineNewspapers.com. Several large directories focus on individual countries or languages, like the Library of Congress' accessible listing of US Newspapers in American Libraries. The Media Ownership Monitor surveys popular media outlets in several countries, but its focus is on ownership, influence, and analysis of countries' media, rather than collecting stories from those publications. The Euromedia Ownership Monitor is a European Union-funded transparency project which provides a searchable database of media sources. Finally, several projects undertake one-time compilations of news sources, often focused on a narrow geographic area and released as part of an ongoing project like the Local News Impact Consortium (Lawrence 2025), or accompanying a publication like the date-bound NELA-GT datasets (Gruppi, Horne, and Adalı 2023).


Figure 3: Story volume over time by publication date.

The primary differences between the Media Cloud Directory and these source lists, apart from its integration with the Story Index and querying system, is its human-curated flexibility and modes of access. Media Cloud hosts many collections geared towards common research needs, includes data about the health of those collections, accepts suggestions of sources to start ingesting, and regularly creates new collections based on both research and projects like those above. Media Cloud is also not limited to a specific geography, nor is it proscriptive in the aspect of news that can be studied (e.g. media bias, media ownership). We also provide API-based access and CSV downloads throughout our system, unlike many of these web-based directories that offer unstructured HTML lists split across a hyperlinked set of pages.

### 4.2 News Content Databases

Commercial options do exist for accessing and studying news content, perhaps chief among them LexisNexis, which archives thousands of sources and has coverage that extends

| Attribute | Description | Notes |
|---|---|---|
| Id | A unique identifier within our system. | A sha256 hash of the normalized story URL. |
| Article Title | Extracted article title. | Based on a hand-curated set of heuristics that search HTML metadata tags and content element tags. Full-text searchable. |
| Canonical Domain | The (almost) unique root domain and suffix of the site that the story came from. | Based on a heuristic method that accommodates country-level domain suffixes and a small handful of custom exception cases. |
| Indexed Date | Timestamp of when the article was indexed. | Generated automatically to track the time the story was added to the Story Index Elasticsearch database. |
| Publication Date | Extracted date of publication. | Based on a tuned comparison between open-source data guesser software libraries. |
| Language | The language most often detected on related text (2-letter ISO 639-1). | Based on a heuristic that tries to guess between any HTML metadata and then falls back to guessing based on the open-source `py3langid` package's evaluation of text content. |
| Text Content | Full text content of the article (not available for download). | Extracted via an ordered list of 3rd party open source packages, created based on internal performance evaluations. Full-text searchable. |
| URL | Full URL of the article. | Based on the final resolved URL after redirects, with common tracking URL parameters removed. Wildcard searchable. |

Table 3: Metadata collected per story in the Story Index.

| Language | Percent of Stories |
|---|---|
| English | 44.4% |
| Spanish | 14.5% |
| German | 7.7% |
| Russian | 6.4% |
| Portuguese | 5.6% |
| Arabic | 4.8% |
| Italian | 4.5% |
| Japanese | 4.2% |
| French | 4.2% |
| Chinese | 3.2% |

Table 4: Distribution of stories in the Story Index by language, for those with at least 2% representation.

further back than Media Cloud, allowing researchers to examine decades worth of coverage (van der Meer, Kroon, and Vliegenthart 2022). However, features such as word frequencies and retrievable results are limited, though the company has an additional analytical suite called Newsdesk that can augment news data. There are also several commercial news APIs like NewsAPI.org, GNews.io, NewsCatcher, Mediastack, NewsAPI.ai, and Newsdata.io. These provide access to up to thousands of sources for a fee, catering to corporate developers monitoring brands rather than researchers (although some have been used in studies). They are generally stronger than Media Cloud for real-time feeds and up-to-the-minute headlines, but less-well suited for any study using historical data, focused on less-popular geographic areas, or requiring flexible, transparently documented collections. They are also primary architected as data services rather than research platforms, although some, like Newsdata.io, provide features like sentiment analysis and crisis detection. Finally, there are scan-based news archives like Newspapers.com, invaluable resources for research that involves pre-web news, but limited by the imprecision of optical character recognition and the difficulty of assigning metadata to physical objects. These are also not suited to conducting bulk text analyses, and as such are best suited for qualitative work. In contrast to all of these commercial databases, Media Cloud is free, open source, and built for studying media ecosystems on the open web, including more than just formal sources.

### 4.3 Event Databases

Whereas the fundamental unit of analysis for a news database is a story or article document, event databases are oriented around events extracted from those stories. Among event databases, the Global Database of Events, Language and Tone (GDELT) is most similar to Media Cloud in that it caters to academic users and functions not just as a data source but as a research platform (Leetaru and Schrodt 2013). GDELT is a free academic project that provides extracted event-related metadata from stories on the open-web. GDELT struggles with duplication and misclassification, however, and has limitations inherent in its dictionary-based pattern matching (Hong et al. 2025). The Integrated Crisis Early Warning System (ICEWS) is another high-profile event database built by Lockheed Martin, intended for use by the US government and not directly accessible (Ward et al. 2013; Wang et al. 2016). Even more focused offering is the Armed Conflict Location and Event Data project (ACLED), with event data on political violence and protests, analyzed by humans rather than just using automated NLP techniques (Raleigh, Kishi, and Linke 2023). For research in its subject area, its freely accessible datasets are important resources for news-derived events. However, ACLED limits public access to the most recent data. Event databases are useful for studying how events evolve, especially inter-

national or multilingual events, while Media Cloud is better suited for researching the language of media and journalism, such as studies of media framing, narrative, choices and journalism itself.

## 5 Research Use

As of December 2025, the Media Cloud front-end analysis tool has nearly 6,000 active users. Journalists, activists, and applied researchers are significant and valued user groups of the tooling. There are a number of notable, socially impactful media analysis projects and tools built on top of Media Cloud's API and data. These include: the Counterdata Network, which draws on Media Cloud data to monitor feminicide in several countries and provide timely data to activists (Bhargava et al. 2022); the Trans News Initiative, which draws from Media Cloud to visualize topic attention and media partisanship within US media on trans-related news stories; and the Canadian Centre on Substance Use and Addictions' Substance Use Monitoring Dashboard, which employs LLM-driven summarization of Media Cloud news data to alert users of emerging substance-related public health concerns. Nonetheless, as Media Cloud is engineered to be a research-quality archive, academic work using the tool is an important measure of its overall impact, and more straightforward to scope for a meta-analysis of Media Cloud's academic use and impact. This analysis is detailed below.

### 5.1 Volume of Citation

Through iterative search and alerting from public scholarly archives, we identified 167 academic papers that used Media Cloud as of November 2025. These exclude numerous teaching materials, white papers, student theses, blog posts, and popular press usages of the tool. Similarly, we did not include publications that only made textual reference to Media Cloud without using it for data or analysis. The first paper found was published in 2011 by the project's former principal investigator at Harvard University, Yochai Benkler (Benkler 2012). Concurrent with the public user interface being developed in 2014, four papers using Media Cloud were released that year, and four or less were published each subsequent year through 2018. During that time period, approximately half of the papers published were authored by internal Media Cloud researchers, and half were authored by external researchers accessing the tooling.

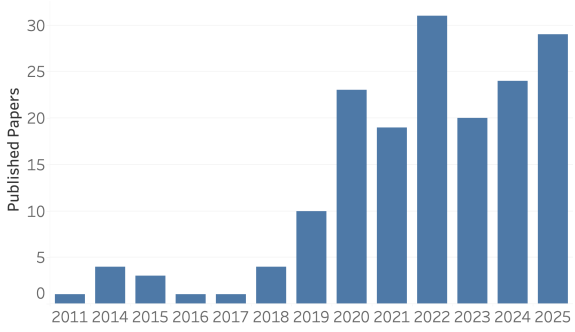


Figure 4: Count of research papers using Media Cloud by publishing year.

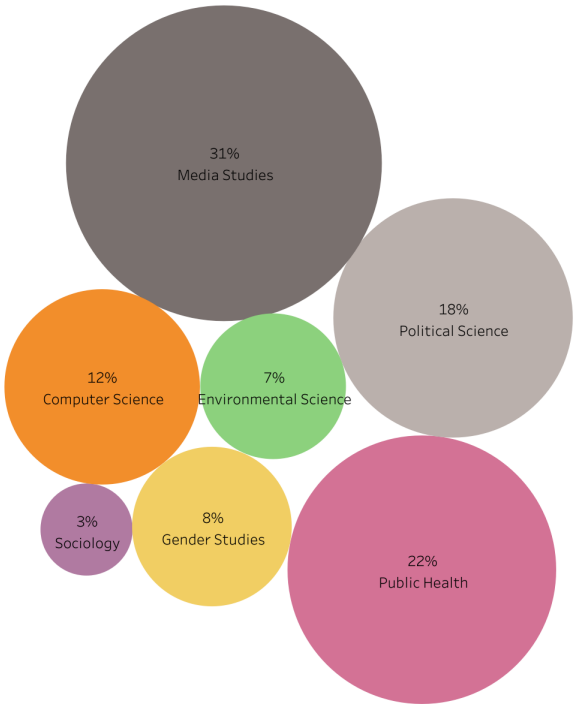


Figure 5: Most prominent primary topics of papers citing Media Cloud.

In 2019 the number of papers published increased to 10, and in 2020 the number more than doubled, with 24 published papers using Media Cloud that year. This spike in usage (Figure 4) was concurrent with the COVID-19 pandemic, a clear application of the Media Cloud tools to research about public health and media coverage.

### 5.2 Fields of Study

To explore topic distribution in a non-rigorous, yet representative way, we employed a chat-based generative LLM to assign three topic labels to each paper. We manually reviewed each result and selected the best two of the three suggested labels, refining where needed (e.g., combining the often duplicative "Journalism Studies" and "Communication Studies" into "Media Studies"), and confirming the correct ordering of primary and secondary topics. This process yielded 28 unique topics of focus within the academic paper corpus (Figure 5).

As expected, media studies is the leading primary topic, representing approximately one-third of all papers. Public health is the second most prominent topic, with slightly more than one in five of all papers having it as a primary topic. Political science is third, at nearly one in five focused on the topic. These are followed by computer science, gender studies, environmental science, and lastly, sociology. When looking at the breadth of all topic labels applied, for both primary and secondary topics (Table 5), again media

| Papers | Topics within Range (descending order) |
|---|---|
| 100+ | Media Studies |
| 10-99 | Public Health, Political Science, Computer Science, Sociology, Gender Studies, Environmental Science, Information Science |
| 2-10 | Data Science, Economics, Latin American Studies, Psychology, Science and Technology, Criminology, International Relations, Race & Ethnicity Studies, Africana Studies, LGBTQ+ Studies, Linguistics, Youth Studies |
| 1 | Asian Studies, Computational Social Science, Food Studies, History, Library Science, Marketing/Business, Peace & Conflict Studies, Tourism Studies |

Table 5: Distribution of papers receiving topic label, as either primary or secondary topic.

studies is the most frequently applicable topic, followed by public health, political science, and computer science.

# 6 Discussion

This paper describes a revised and updated Media Cloud system that offers two central datasets for studying news online: a Directory of news sources and a Story Index of online news published by those sources. These datasets are the latest evolution of the multi-decade Media Cloud project, including more extensive source de-duplication and harmonization of metadata across stories. These datasets are already driving broad types of study across a wide diversity of research domains, and offer opportunities for many others.

## 6.1 Limitations

Users of the datasets must be aware that global digital news production, which is what Media Cloud aims to capture, is a complex and constantly evolving phenomenon. Regardless of how we define news, there are many thousands of digital outlets providing information on current affairs in hundreds of languages around the world. Simply identifying those sources is a huge challenge, and we are aware that our collections have gaps.

Acquiring and storing the content produced by sources is also a complex and time-sensitive matter. We aim to collect all content and do so as soon as possible to allow research on current affairs, and before it disappears. However it is impossible to collect all the content from all sources, because there are paywalls and blocks, because the content is published in formats that cannot be easily captured by our tools, because things break and data flows are interrupted. We aim for comprehensiveness while documenting our gaps. These points of friction have been increasing over the lifetime of our project, likely driven by both business (such as paywalls for revenue) and technology aspects (such as scraping for LLMs). As one response to the observed decline in RSS feeds as a widespread mode of syndication, this new system adds support for Google News Sitemaps (Google News 2010).

Our goal to capture as much data as possible also involves tradeoffs. The legacy system was very rich in metadata and data pre-processing, which made it very hard to maintain once the database passed a certain size. The new system, in contrast, provides less metadata but is easier to maintain, sustain, and scale. I addition, the progressive closure of social media platforms' APIs makes it very difficult to understand the digital environment as an ecosystem, and to trace the circulation and impact of news beyond the open web. Another limitation relates to copyright restrictions on sharing full-text of articles, which inhibits certain types of research and often forces users to fetch content for themselves.

## 6.2 Conclusion

Despite these limitations, Media Cloud continues to serve as an important reference for the digital research community after almost two decades. Our analysis of academic citations shows it has established itself as a key piece of digital public infrastructure (including being awarded that designation by the UNDP) whose breadth and importance to scientific research has only increased over the years. We are building new ways to interact with our collections of sources so they can be updated to remain current and relevant to the media ecosystems they represent, we are making constant improvements to our back-end to streamline it and make it easy to monitor and maintain, and we are building new analytical and research tools to compensate for the limited metadata the systems offers by default.

This work exists against a backdrop of fundamentally altered research capacities for media scholars. Media Cloud started at a time when the open web and its hyperlinked structure constituted the constantly expanding core of the digital public sphere, a time when social media had not achieved prominence, and when generative AI seemed but a distant dream. Despite the growing impact of web-based media on broader parts of society, dominant models of data sharing have been eroded in ways that make research uses challenging to impossible. In this "post-API age" it is no longer possible to pull rich data from Facebook, X/Twitter, or other social media platforms (Freelon 2018). Media Cloud remains a defiant exception, providing the equivalent of an API to study journalistic media with modern quantitative methods.

The system and datasets have evolved, and continues to evolve, in parallel to the changes in the digital environment and the new challenges they bring. Media Cloud remains a valuable piece of digital infrastructure to researchers, and continues to strive for openness, comprehensiveness, sustainability, and rigor in the datasets we provide.

# Acknowledgements

The bulk of this work on Media Cloud tooling and datasets was supported by National Science Foundation grant #2341858. We'd also like to thank our past and present funding partners who have made this work possible, including the Gates Foundation, the Ford Foundation, the Knight Foundation, the Robert Wood Johnson Foundation, and the MacArthur Foundation. Thank you also to long-time leaders

and champions of the project who have moved on: Yochai Benkler, Hal Roberts and Rob Faris. Finally, we extend our gratitude to the numerous contributors to this open source project over its more than a decade of existence, including researchers, developers, designers, journalists, librarians, students, and our global user community.

## References

Barbaresi, A. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 122–131. Association for Computational Linguistics.

Benkler, Y. 2012. A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate. *Politik*, 15(2).

Bhargava, R.; Suresh, H.; Dogan, A. L.; So, W.; Suárez Val, H.; Fumega, S.; and D'Ignazio, C. 2022. News as Data for Activists: a case study in feminicide counterdata production. In *Computation + Journalism*.

Borah, P.; Ghosh, S.; Hwang, J.; Shah, D. V.; and Brauer, M. 2024. Red Media vs. Blue Media: Social Distancing and Partisan News Media Use during the COVID-19 Pandemic. *Health Communication*, 39(2): 417–427. Publisher: Routledge _eprint: https://doi.org/10.1080/10410236.2023.2167584.

Chen, K.; Babaeianjelodar, M.; Shi, Y.; Janmohamed, K.; Sarkar, R.; Weber, I.; Davidson, T.; Choudhury, M. D.; Huang, J.; Yadav, S.; Khudabukhsh, A.; Nakov, P. I.; Bauch, C.; Papakyriakopoulos, O.; Khoshnood, K.; and Kumar, N. 2022. Partisan US News Media Representations of Syrian Refugees. arXiv:2206.09024 [cs].

Dhillon, P. S.; Panda, A.; and Hemphill, L. 2025. How digital paywalls shape news coverage. *PNAS Nexus*, 4(1): pgae511.

Elfes, J. 2025. Mapping News Narratives Using LLMs and Narrative-Structured Text Embeddings. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, 326–337. Association for Computing Machinery. ISBN 979-8-4007-1483-2.

Fletcher, R.; and Nielsen, R. K. 2017. Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication*, 67(4): 476–498.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Freelon, D. 2018. Computational Research in the Post-API Age. *Political Communication*, 35(4): 665–668. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2018.1477506.

Google News. 2010. Google News Sitemaps. https://www.google.com/schemas/sitemap-news/0.9/sitemap-news.xsd.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2023. NELA-GT-2022: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. arXiv:2203.05659 [cs].

Hong, D.; Fu, Z.; Zhang, X.; and Pan, Y. 2025. Research on the Development and Application of the GDELT Event Database. *Data*, 10(10): 158. Publisher: Multidisciplinary Digital Publishing Institute.

International Organization for Standardization. 2017. ISO 28500: WARC file format. https://www.iso.org/standard/68004.html.

Kavanagh, J.; and Rich, M. 2018. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. RAND Corporation. ISBN 978-0-8330-9874-0.

Lawrence, R. 2025. Newsroom Census/Ecosystem Mapping Toolkit. https://www.localnewsimpact.org/wp-content/uploads/2025/11/LNIC-Toolkit-V1.pdf.

Leetaru, K.; and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, 1–49. Citeseer.

Qian, S.; Chen, K.; Meng, J.; Shen, C.; Chen, A.; and Zhang, J. 2025. Fear in Media Headlines Increases Public Risk Perceptions but Decreases Preventive Behaviors: A Multi-Country Study During the COVID-19 Pandemic. *Journal of Health Communication*, 30(1): 29–39. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10810730.2024.2439468.

Raleigh, C.; Kishi, R.; and Linke, A. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10(1): 74. Publisher: Palgrave.

Roberts, H.; Bhargava, R.; Valiukas, L.; Jen, D.; Malik, M. M.; Bishop, C. S.; Ndulue, E. B.; Dave, A.; Clark, J.; Etling, B.; Faris, R.; Shah, A.; Rubinovitz, J.; Hope, A.; D'Ignazio, C.; Bermejo, F.; Benkler, Y.; and Zuckerman, E. 2021. Media Cloud: Massive Open Source Collection of Global News on the Open Web. *Proceedings of the International AAAI Conference on Web and Social Media*, 15: 1034–1045.

St. Aubin, C.; and Liedke, J. 2025. Social Media and News Fact Sheet - Pew Research. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

van der Meer, T. G. L. A.; Kroon, A. C.; and Vliegenthart, R. 2022. Do News Media Kill? How a Biased News Reality can Overshadow Real Societal Risks, The Case of Aviation and Road Traffic Accidents. *Social Forces*, 101(1): 506–530.

Wang, W.; Kennedy, R.; Lazer, D.; and Ramakrishnan, N. 2016. Growing pains for global monitoring of societal events. *Science*, 353: 1502–1503.

Ward, M.; Beger, A.; Cutler, J.; Dickenson, M.; Dorff, C.; and Radford, B. 2013. Comparing GDELT and ICEWS event data. *Analysis*, 21: 267–297.

Zuckerman, E. 2004. Global Attention Profiles - a Working Paper: First Steps Towards a Quantitative Approach to the Study of Media Attention. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=487943.

## Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, due to respectful fetching rate of fetches for web content, ability for sources to be removed from the Story Index, and limitations on sharing copyrighted content
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, this is a dataset paper and we document the dataset and architecture
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? NA
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, including notes about over-representation in the data based on the authors' topics and geographic regions of study
   (e) Did you describe the limitations of your work? Yes, in the section 6.1
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in various parts of section 2
   (g) Did you discuss any potential misuse of your work? Yes, in various parts of section 2
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in various parts of section 2
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we rely on a number of open-source packages and mention them.
   (b) Did you mention the license of the assets? No, because the software isn't part of the dataset we felt that was out of scope
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we will post the updated Geographic Collections dataset to a data repository if the paper is accepted.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we mention copyright concerns
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? This is general news published on the web so we aren't releasing anything not already available
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR ? Yes, we have a section on this in section 2.3
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No, because the system documented allow for users to create their own datasets
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA