9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

The News Observatory: A Collection and Storage System for Past and Future News Media Content

Keywords: Software, News, Data Collection, Reproducibility, Infrastructure

Extended Abstract

In the 21st century, news agencies have been moving their business onto the internet. Universities and other research organizations are increasingly interested in tracking the behavior of news agencies. Since news agencies produce large sums of text data every day and distribute it as HTML, the process of gathering and collating a sufficiently dense dataset of news articles is technically difficult, especially if we also wish to extract structural information from the HTML. News websites are also inherently fast-moving; while most are static https servers, the front page changes rapidly because the content must be "current".

The ephemeral nature of the internet poses a challenge to researchers. In order to analyze historical trends, we need a good source of historical data. The Internet Archive's Wayback Machine serves as a good source of historical html documents, but has structurally sacrificed specificity for scale [Archive.org, 2023]. Volunteers upload documents to the Wayback Machine at their own discretion, and any and all web documents are accepted. In order to support the wide scope of the Wayback Machine, only raw html is collected, which puts the impetus for processing on the researcher.

We present a system for automatically collecting and parsing Wayback Machine entries into a dataset tailored for exploring news publisher behavior on the internet. The system can be broken up into three subsystems:

First, the scanning subsystem: a complete list of Wayback Machine entries is retrieved from the Internet Archive. Entries with non-200 status codes are filtered out, and we process the remaining pages to generate a sufficiently dense list of entries. This process works by generating a list of "ideal" timestamps (for the News Observatory this is hourly during daylight hours in the United States), then identifying real entries with timestamps closest to the ideal time without being before. The core of the scanning subsystem has been released as a standalone command line tool.

Second, there is a parsing subsystem. The parsing subsystem takes every entry identified by the scanning subsystem, and parses the raw html to infer a "ranking" of articles. Parsing behavior is tailored to each publisher, but generally articles are considered more important if they appear higher on the page, further to the left, are associated with larger images, and whose headlines are rendered with a larger font. The process of tailoring the parser to a publisher is semi-automated and modular. The list of articles is then passed to the collector, the final subsystem.

Using information derived from the parser, the collector will ask the scanner for the nearest article entry to the associated front page entry (without going backwards). Each of these article entries is converted to plain text using Trafilatura [Barbaresi, 2021] and Readability [Mozilla, 2023] and stored in a relational database. Every article entry is indexed with a checksum, to ensure that duplicate entries are properly reported and stored efficiently. The relational database is designed such that a user may explore the dataset from a number of different perspectives. For example, the database can be scanned in terms of changes to the text of an article,

to see how an article or class of articles tends to be edited over the course of its lifespan. A user could also scan the database with respect to front pages, unifying each article in terms of its location in the site map, to analyze how article publication has changed over time.

These subsystems are then integrated into a cloud computing environment for automation and to maintain long-term data health. Microservices that periodically update the database with new articles, check for inconsistencies in articles, and otherwise provide Continuous Integration are run on timers and triggers to minimize maintenance costs, allowing us to host the News Observatory for years to come.

Data exploration tools like this have been built in the past, particular examples include Media Cloud [Roberts et al., 2021] and Lexis Nexis [RELX, 1970]. Lexis Nexis presents itself as a general purpose database that happens to index media. Media Cloud presents itself as an open source tool for analyzing media behavior online, and is also based off of Internet Archive data. Both of these tools focus on generating a wide dataset, and providing summary data or a selective subset of data to users. We recognize the achievements of these systems and others, as they have led to meaningful and interesting discoveries. In contrast, the News Observatory generates a more focused dataset, currently focused on mainstream media outlets. This allows us to tailor our processing systems to the unique challenges of studying online news. For example, the News Observatory provides information on article placement within the front page, differential changes in article text over time, and hourly changes to headlines and placements. In addition, the subsystem-focused design of the News Observatory allows for simple development of new metadata objects in the future. Each subsystem is designed as modularly as possible, for example the parsing subsystem can be extended to other publishers by simply providing high-level structural information. This structure will allow us to strategically mine the Wayback Machine for project-specific requirements, giving researchers greater exploratory freedom in the future.

We have already found the News Observatory database to be useful in our ongoing research, and it is our hope that it can be used to accelerate computational media research at large.

References

[Archive.org, 2023] Archive.org (1996-2023). The wayback machine.

[Barbaresi, 2021] 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*, pages 122–131. Association for Computational Linguistics.

[Mozilla, 2023] Mozilla (2023). Readability.js. Mozilla.

[RELX, 1970] RELX (1970). Nexis: Online Business Research & News Media Database — Nexis. https://www.lexisnexis.com/en-us/professional/nexis/nexis.page.

[Roberts et al., 2021] 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(1):1034–1045.

Figures

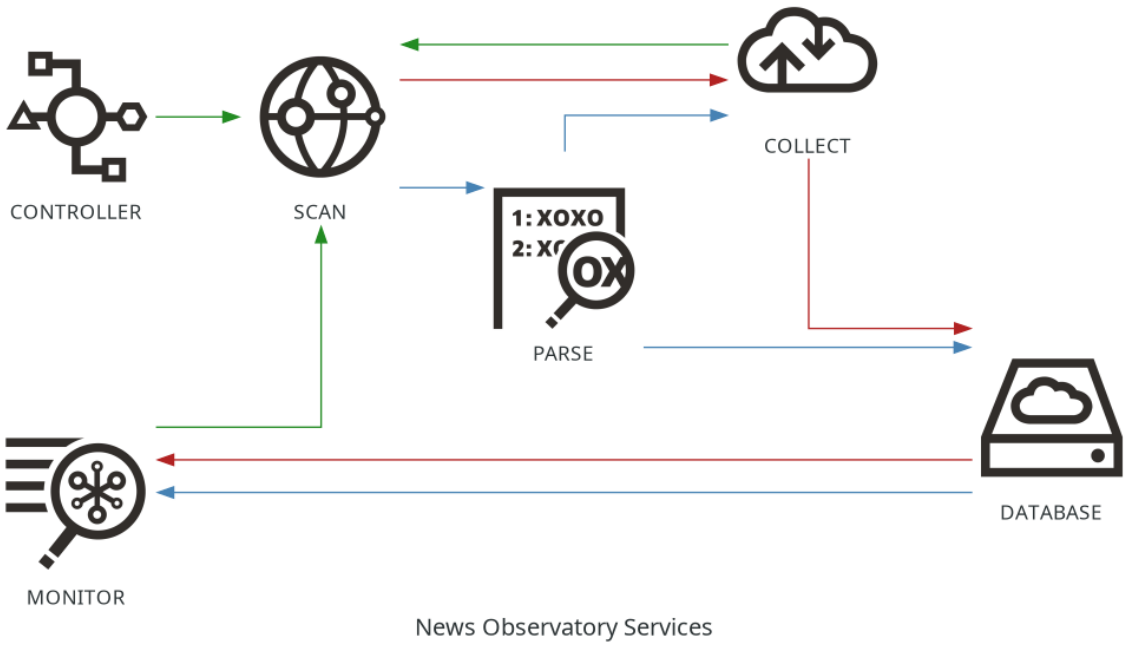

Figure 1: A diagram of the services and subsystems that make up the News Observatory. The front page collection cycle is shown in blue. The article collection cycle is shown in red. Control signals are shown in green. ”MONITOR” and ”CONTROLLER” represent two groups of microservices that independently and dependently trigger actions in the rest of the system.