Efficient and reliable utilization of automated data collection applied to news on climate change

Erkki Mervaala and **Jari Lyytimäki**

Finnish Environment Institute Syke, Finland
firstname.lastname@syke.fi

Abstract

Automated data collection provides tempting opportunities for social sciences and humanities studies. Abundant data accumulating in various digital archives allows more comprehensive, timely and cost-efficient ways of harvesting and processing information. While easing or even removing some of the key problems, such as laborious and time-consuming data collection and potential errors and biases related to subjective coding of materials and distortions caused by focus on small samples, automated methods also bring in new risks such as poor understanding of contexts of the data or non-recognition of underlying systematic errors or missing information. Results from testing different methods to collect data describing newspaper coverage of climate change in Finland emphasize that fully relying on automatable tools such as media scrapers has its limitations and can provide comprehensive but incomplete document acquisition for research. Many of these limitations can, however, be addressed and not all of them rely on manual control.

1 Introduction

Despite the digital era's advancements, manual data collection continues to dominate humanities and social science studies, notably in media studies where the significance of digital communication is ever-increasing (Shearer and Mitchell 2021). Most print newspapers publish also online versions of their news content, and these online versions have exhibited modest variations in content compared to their print counterparts (Hoffman 2006; Mensing and Greer 2013; Hagar and Diakopoulos 2019).

The growth of online data has spurred the development of various automated data collection tools, such as media scrapers and public APIs, enhancing accessibility to vast datasets (Sirisuriya 2015; Aitamurto and Lewis 2013). However, the ease of collecting big data has potentially overshadowed inherent biases and errors, leading to availability bias and other types of bias affecting dataset representativeness (Mahrt and Scharkow 2013; Grimmer et al. 2022).

While web scraping is often viewed as a technical phenomenon, there is a growing discourse on the "softer issues" surrounding it, including ethical and legal considerations (Murray State University et al. 2020; Khder 2021; Zimmer 2010; Bruns 2019). The field is evolving, especially as platforms like Meta and Twitter have restricted data access.

Research on automated data collection has proliferated since the turn of the millennium, focusing largely on social media content (Scharkow 2013; Venturini and Rogers 2019). However, less attention has been given to utilizing automated methods for newspapers, with warnings about the trade-offs between automation and reliability (Deacon 2007; Mahrt and Scharkow 2013; Wijfjes 2017).

Media content analysis has traditionally involved small samples and qualitative approaches due to labor-intensive collection and coding. The shift towards automated research methods is motivated by the potential for larger sample sizes, despite reliability trade-offs (Broersma and Harbers 2018; De Grove et al. 2020; Wijfjes 2017; Blatchford 2020). Challenges and caveats related to computational methods, including supervised machine learning, have been discussed, emphasizing the need for caution in overestimating the benefits of automation (De Grove et al. 2020).

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 82–91
December 1-3, 2023 ©2023 Association for Computational Linguistics

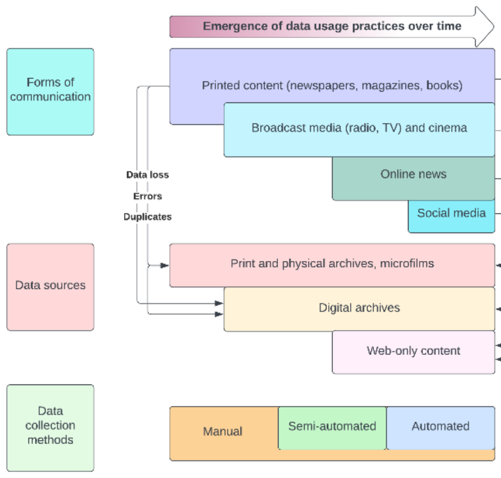

Figure 1: Evolution of data usage for media studies. The figure expresses data sources and usages of different media.

Media studies often lean towards manual or semi-automated collection methods, with less emphasis on fully-automated tools or "theory-driven online scraping" (Lodhia 2010; Khder 2021).

In Figure 1, we summarize the evolution of data usage and data collection methods and issues related to the reliability of data archiving from platform to platform. The aim of this article is to critically examine the pros and cons of different data collection methods and the crossing from manual and semi-automated data collection to fully automated practices. It is based on a case study focusing on newspaper data on climate change, showing the development of climate change news from 1990 up to December 2020.

2 Methods and materials

Our focus is on the news coverage of climate change in the Finnish newspaper Helsingin Sanomat (HS), given its high societal relevance, interdisciplinary character, and extensive previous studies on its climate coverage (Suhonen 1994, Lyytimäki 2011, Kumpu 2016, Teräväinen et al. 2011, Ylä-Anttila et al. 2018, Boykoff et al. 2019, Lyytimäki 2020). HS, the most widely circulated newspaper in the Nordic countries, has been a key source for monitoring media coverage of climate change in 58 countries and is a common subject in digital humanities and media studies (Boykoff et al. 2022).

The manual data (MD) for comparison comprises 14,750 news stories headlines retrieved from HS's online archive, spanning from January 1st, 1990, to December 31st, 2020. These stories, collected into a spreadsheet, were identified using specific climate-related queries (search screening full texts and using Finnish search terms for climate change, warming of climate and greenhouse effect) and included even those items mentioning climate issues tangentially (Lyytimäki 2011, 2015, Lyytimäki et al. 2020). Duplicates and irrelevant hits were removed based on manual inspection. Various factors, such as changes in the newspaper structure and search engine properties, influenced the data's format and content, with different information available across years and some data, like cartoons and advertisements, excluded.

Automated data were obtained using two different scrapers utilizing the Sanoma API. The first scraper (S1) mimicked the manual approach, collecting data in batches of 50 articles, mimicking the batch size of articles the manual online search provides after each click of the "show more" button, from oldest to newest, including full texts where possible, using the newspaper3k Python package. The second scraper (S2), based on the Finnish Media Scrapers project (Mäkelä and Toivanen 2021), performed 93 queries to the API, breaking down the search period into weekly segments and yearly intervals for each query term. As the manual dataset consisted only of headlines, publication dates and the article urls, the scrapers were set to collect only those data.

Both scraped datasets underwent cleaning to remove exact duplicates and ensure uniform formatting. The final comparison between manual and scraped datasets involved further cleaning and unifying data formats, focusing on the months the articles were published.

It is crucial to recognize that while MD, S1, and S2 all access the same news archive, the methodologies employed by each distinctly shape the dataset's composition. This underlines the significance of the data collection process itself, as it inherently filters and frames the information extracted from the archive. Therefore, any disparities in the collected data are attributed to the differences in collection methods and the inherent biases each method may introduce, rather than variations in the source material except in the cases when changes had been made to the archive's content or categorization in the times between the manual and scraped data collection.

While we acknowledge that inherent differences in the approaches of MD, S1, and S2 methods may lead to variations in the collected data, the comparison aims to highlight the nuances and potential biases each method introduces. The objective is to understand the trade-offs between manual and automated data collection, aiming to highlight the nuanced insights each approach offers and the unique biases they may introduce to the research on newspaper articles.

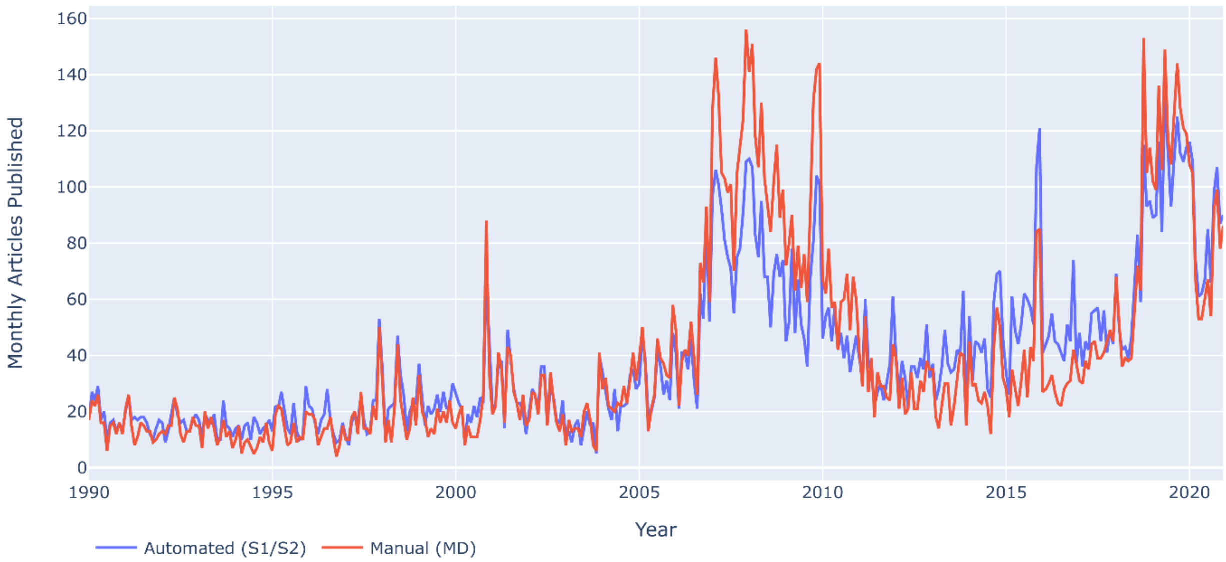

Figure 2: Articles on climate change published on Helsingin Sanomat 2000 – 2021 collected from online archive. The figure shows clear peaks in the frequency of climate change coverage but also highlights differences between the datasets.

3 Results

Compared to the manual dataset (MD) of 14750 news articles, neither of the datasets collected via the automated scrapers gave the exact same result. Also, different scraping techniques resulted in different amounts of articles. ¨

The S1 scraper queries resulted in 8227 stories on climate change, 7441 stories on greenhouse and 1576 on climate warming. After removing duplicates there were 14669 news articles published between January 3rd 1990 and December 31st 2020. The first article of the dataset details record heat in England and the last headline of the dataset declares that the year 2020 was the warmest year on record in Finland.

Initially, the S2 scraper provided the least amount of results: 7970 stories on climate change, 7437 on greenhouse and 1575 on climate warming with a total of 14553 articles after removing duplicates. Representing both retrieval and resource bias (Grimmer et al., 2022), the reason for the scraper collecting fewer articles than the other two is that the scraper ran into problems with either broken articles, manifesting as blank pages or error messages, or articles consisting of dynamic content that prevented scraping the full texts of the articles. After correcting this and limiting the results to article headlines only, S2 resulted in an almost identical result as the first scraper with only one article more, on climate change, than S1. From here on, we will discuss only the S1 dataset.

The full manual dataset of 14750 articles had 81 articles more than the 14669 of S1 (See Fig 1). While the difference between the datasets is only half a per cent in total numbers, the differences become more apparent when comparing certain peaks in the data: In November 2000 S1 dataset showed 69 published articles and MD 88 articles. Other similar peaks include February 2007 (S1: 106, MD: 146) and February 2008 (S1: 109, MD: 156). From 2011 to 2018, the S1 seems to take over and contain more results. The largest peaks of S1 align with the December 2015 Paris Accord when S1 displayed 121 results and MD only 85. From 2018 to the beginning of 2020, MD displays more results on average and after that S1 again until the end of the year 2020.

On closer inspection, including a detailed manual review of the discrepancies, focusing on the type and content of articles that differ between the datasets, the articles causing the differences are mainly smaller commentaries, opinion pieces or editorials, and on a smaller scale, television or radio programming details. For December 2007 MD has 156 articles and S1 had 109 articles. The differences appear to come from more opinion piece articles included in the manual dataset compared to the scraped set. While some opinion pieces and editorials were included in the scraped set, MD included numerous relevant ones such as a small comment piece titled “Vuoden viherpesu” (“Green Wash Of The Year”).

In the opposite case of December 2015, the surplus of articles in the scraped dataset is mainly the result of several different editions of the same story published on two different sections of the site such as “ulkomaat” (“foreign”) and “ilta” (“evening”). In addition, some opinion pieces were included in the scraped set that were not present in the manual set.

When calculating the percentage of matching articles between the datasets, using their unique identifiers, the article headlines and urls, the datasets were only 84,2 % identical. The differences can be mostly explained by differences in coding the articles in the manual set and the automatically retrieved headlines from the online archive which in turn may also change over time especially if the articles were subjected to A/B testing, usually changing the articles’ headlines to optimize online readership, during or after the data collecting. It should be pointed out that in February 2023 an editor of Helsingin Sanomat admitted to modifying headlines of their online and print versions differently and an editor of the evening tabloid Iltalehti stated that negative headlines work better as they interest people more (Sillanmäki, 2023).

These kinds of discrepancies should, however, be also accounted for when assessing different ways of obtaining data. A more reliable way to compare articles would be to use the articles’ hyperlinks that are not likely to change over time.

Considering a stricter approach to removing duplicates, some articles were indeed almost identical to each other when it comes to the headline and even the article content despite having different hyperlinks. Removing duplicates based solely on the title or solely on the hyperlink may still leave different versions of the article in the datasets as some archived articles from the beginning of the datasets’ time period may have both the print version and the online version of the article available online with individual hyperlinks with minor variations in the online headline. In some cases, the same article was published twice within the same month with a different hyperlink. Also, the same or very similar headlines may lead to a “full” and an “abridged” version of the story. A combination of filtering by unique hyperlinks and headlines with the possible addition of publication month and content comparison may be a more accurate, though more cumbersome, approach.

4 Discussion

4.1 Automation as a solution

Updates in search engines and content and categorizations of the database may distort search results updating old data. It is also possible that some items related to climate issues are missing from the sample because of the limited set of keywords. Therefore, it is vital to conduct test searches to ensure that the right balance is found between exclusion and inclusion. This, in turn, requires expertise on the qualities of the issues under scrutiny. For example, coverage of biodiversity loss and “the polycrisis” may overlap with climate change coverage.

While manual data collection can offer a relevancy filter of sorts already during the collecting process, it is slow as all the details of the articles have to be manually copied and pasted or written in the data set document. The manual collecting process raises also issues with repeatability and handling errors in the original tasks found out later during the process. Especially with vast datasets, noticing an error after the data has been collected, it may not be possible to repeat the process afterwards due to limited human resources. The speed of automated data collection depends mainly on the processing power attributed to the scraper and the amounts of articles published during the period in question. For example, scraping article headlines for the search query “climate change” can take anything between a few seconds to a few minutes. For manual collection, the time spent can be considerably longer (Lauer et al. 2018), often beyond the resources available. Although automated scraping significantly enhances cost-efficiency and data breadth, it is not without trade-offs. For instance, automated methods may inadvertently capture irrelevant data, necessitating post-collection filtering that can be both labor-intensive and prone to oversight. This underscores the importance of a balanced approach that weighs the speed and scope of automation against the precision and context sensitivity of manual data collection..

The automated method also offers the possibility to collect much larger datasets much quicker and therefore the possibility of more comprehensive scopes for studies even if the data would have to be filtered down later. Manual collection can also suffer from a lack of timeliness as collecting the data can be too slow to produce data fast enough for topical analysis on quickly evolving topics. Apart from the comparable slowness, additional human errors and biases can be coped with via well-established ways such as intercoder reliability tests.

Relying on automated methods may easily lead to omissions in reliability testing as data collected automatically can be assumed to have been collected “objectively”. In order to find the most reliable solution, testing between different automated methods and comparing results to similarly produced manual samples would be one way to address this issue, albeit time-consuming. The need for such testing increases with the gaps between data collection sessions as changes in APIs may result in different search results.

. Especially with larger datasets consisting of thousands or millions of data points, systematic errors, that might have been caught more easily by human eyes, may go unnoticed by the researcher relying on automated data collection. Therefore, testing the methodology via smaller test runs is encouraged. While a scraper can perform perfectly fine for 90 per cent of the news articles, the remaining ten per cent may cause issues for the whole dataset. For example, a single misplaced comma or a semicolon scraped in the scraped data may mess up the following rows and columns. Additionally, especially on archived content, the scraper may hit a wall due to bad or obsolete programming. Such issues arise most often when scraping for full articles as each news story is a page of its own for the scraper to run into error-inducing content which at best may lead to empty content cells in the dataset. For these reasons, error handling is very important in the scraping process.

Causes for such systemic errors can also change over time. For example, changes in the newspaper website infrastructure such as adding CAPTCHA, a program that checks whether the user is human or a machine, and other anti-scraper measures will affect the results and possibly prevent for example collecting full texts of articles especially if the articles themselves are behind a paywall. Additionally, the introduction of the so-called “dynamic articles” that feature semi-interactive and interactive elements that reveal text as the reader scrolls down the article, also affects collecting the full texts of the articles, as they often require more sophisticated scraping techniques, frequently requiring site-specific programming. Such dynamic articles may be challenging for manual data collection as well.

Finally, there are possible issues with timestamping the data. As the data is for the exact times when the articles were published and modified are available via scraping, there is a need to normalize the ordering of the data in the dataset whether it be by year, month, day or by minute. Whereas in fast-paced social media communication it may be important to know the publishing time by the second, in online news media analysis the timestamping may not need to be as detailed. The article can also be modified or republished after its original publication which may lead to the article being misplaced in the dataset depending on which variable one uses to sort articles by – for example “time published” or “time modified”. Though an issue of potentially limited relevance, should an article be updated for instance at the change of a month, it may be duplicated in a collection of datasets updated monthly. Additionally, the order of the articles may be relevant for consequential articles covering short-lived, fast-paced events.

4.2 Common challenges

There are also several common challenges for both manual and automated data collection. Changes in visual design and composition of the sections of the newspaper may have an influence on the number, length, and presentation style of news items. For example, during the study period, the composition of the printed version of HS was renewed several times, including a major change from broadsheet to tabloid on 8 January 2013. (Sanoma 2012). The data itself may not be complete as the provider may have altered the archive over the years. These kinds of archive alterations may not have had any nefarious intentions behind them as they may have been part of restructuring the archive for better accessibility or functionality and may be limited to actions such as removing duplicates or recategorizing content. In some cases in the HS dataset, duplicate versions of articles were found even with a different hyperlink as they represented different versions such as online and print versions of the same article with only minimal changes.

Proper (automated) comparison of the manual and scraped datasets require some unification and cleaning for the data. As the manual collecting process for large datasets often includes more than one researcher and may stretch to long periods of time, differences in recording the data are bound to be more frequent compared to automated scrapers that perform the task without variations. Omitted details can for example be added to the manual datasets using even the same automated tools used for scraping. It should be noted that each comparison case is different, and the methods and tools required to address such issues should be assessed by case and by data type.

The transformation of news media from static text to dynamic, multimedia narratives presents both opportunities and challenges for data collection. Visual elements like photographs, infographics, and videos are integral to modern storytelling and can significantly influence audience perception. However, these non-textual elements are often not captured by traditional scraping techniques, highlighting a gap in our methodology that future studies will need to bridge to fully understand media impact.. Additionally, in recent years we have seen an uptick in different kinds of more complex news content such as the aforementioned dynamic news articles, and interactive news articles with sliders, polls and calculators, both providing valuable journalistic content and even significant amounts of text data to the reader but more complex to include as part of a text-based study. Embedded content may also prove to be difficult to access in the future, especially if it is included content that has since been deleted from the source. Deleted Tweets from Twitter/X, for example, are not accessible via those news articles that have embedded them in the middle of the news text after the deletion. Even though the contents of such Tweets would have been written out within the news text, they often are not verbatim and, if not in the native language of the publication, are translated.

These issues reflect the overall evolution of a news article and the structural changes of news over time. Are both a long-form written piece and a news item including infographics and info boxes considered individual news stories? What about stories that are ever-changing or constantly updated such as articles following the global carbon budget diminishing every minute or articles related to the COVID-19 pandemic with daily updates on infections and victims? One way to individualize an article could be based on the article's hyperlink. Then, if the article is changed, the hyperlink stays the same. This, however, does not take into account the potential changes in the message the article conveys to the reader. An article's headline can change several times during the day of the publication due to click optimization, A-B testing, and localization to name a few reasons (Hagar and Diakopoulos 2019). The "original" headline could be said to be the one appearing on the paper version of the newspaper but then articles without a printed counterpart would have to be omitted.

It is therefore paramount for the transparency and reproducibility of the data that a timestamp of the data collection is included also in the dataset. As changes and corrections in the text are often highlighted in the articles in question after the fact, the timestamp, while not covering the change, can at least indicate whether the article was included in the dataset before or after the alteration.

The issue with the changing headlines is a recent one but an important one. While we do not focus on the messaging and framings in the headline in this article, the changes made to headlines that appear to the readers in different forms over different times, devices and platforms is an important topic for media studies and would have to involve tools closely monitoring such changes. A similar approach could and should be applied to the changes in the content of the articles. In fact, there are some instances that already collect and publish changes in headlines and content of news publications online.[i]

4.3 Editorial decisions and the evolution of the language used

The caveats for any use of automated online search functions of newspapers include the possibility that there may be articles omitted from the dataset that could be argued to be categorized as related to a topic such as "climate change" but for some reason have not been included. These omissions could, however, be argued to represent in a rather transparent way the views of the news outlets. If an article is not included in the search results, whether on purpose or not, the media outlets communicate to their readers that the article in question is in fact not relevant in that context. The lack of categorization of the "missing articles" may, of course, have other, "human" reasons, too. The time and resource constraints at the media organization may play a role, as well as potentially the expertise dealing with the categorization, especially if done manually, may lead to the omission of some articles appearing relevant to climate scientists but perhaps not to the media in question. The primary category attached to the article may also be a factor, as several crises such as food shortages may in fact have to do with climate change but are not categorized primarily as such.

The historical topic relevancy is also a factor, and search strategies should allow comparisons between different times and places. Climate change provides an example of a global issue with shared key terminology across different contexts, but languages differ in their emphasis as exemplified by the lack of use of the term "global warming" in Finnish debate. The language used to describe climate change has evolved considerably over the years, which is apparent in the data as we look at the yearly datasets by the scraper search queries: in 1990 there were 18 articles categorized as "climate change", 16 articles as "climate warming", and 295 articles on "greenhouse*", respectively, while in 2020 the respective figures were 1052, 82, and 288. Not only did the amount of the articles increase but also the shift to using the term "climate change" ("ilmastonmuutos") instead of "greenhouse effect" ("kasvihuoneilmiö") is apparent. By sheer quantity, the switch seems to have happened between 2006 and 2007, which coincides with the publication of the influential Stern Review on the Economics of Climate Change (Stern 2007) released in October 2006.

Additionally, even if the news story on climate change has been categorized by a news outlet in the category "climate change", the article may still be omitted from search results with the search query "climate change" for some other reason unknown to the public. For example, recent climate coverage in Finland often deals with carbon sinks of the Finnish forestry not necessarily mentioning the term climate change and labelled under energy policy rather than climate policy. The same retrieval bias applies to the concept of "emissions" as relevant stories may include references to emission targets but not climate change specifically. Furthermore, the apparent easiness of using such digital databases may tempt simplification in framing a complex topic such as climate change and prompt conclusions omitting the context. Similar simplification has been found for example in the coverage of Africa (Madrid-Morales 2020).

All in all, the Finnish newspaper archiving system does offer a wide array of opportunities for research: Historical newspapers are comprehensively digitalized with public and free access as their copyrights have already expired. While there are no comprehensive digital archives for more recent media coverage, the consolidation of media companies has led to archives combining materials from some previously independent newspapers. In these cases, the availability of copyrighted materials depends on the right owner. Access to such easy-to-use digital archives may also limit the usage of a certain database over another. HS not only provides the digital archive from 1990 onwards but also an archive of digital replicas of their newspapers from 1889 to 1997 in PDF format. Full texts are made available for subscribers. The PDF archive is, however, not as easy to analyze via automation and machine learning and would require for example tools related to computer vision.

Finally, compared to research on print editions or their virtual counterparts such as PDF copies, online news archives are unable to provide information on the visibility given to the article on the day of publication. Though the front page of the print edition and the main stories on the web page do frequently differ, online news archives only tell when the story was been published with possible additions of its categorization and type.

5 Conclusion

Our findings reveal the impracticality of an exhaustive data collection strategy, challenging the notion that completeness equates to comprehensiveness. Instead, our research underscores the need for strategic sampling, where the focus is on capturing a representative swath of articles that collectively provide insight into the evolution and nuances of issues such as climate change coverage. Whether collected via automated scrapers or manual methods, it is very likely that all the news articles published will not be included in the dataset. There is a risk of complete lack and omissions of data for poorly deposited early years and risks related to diversifying presentation formats for recent years. Significant caveats should be addressed remaining caveats always communicated effectively.

In order to avoid the research methodology becoming a black box, we advocate for meticulous documentation of data collection processes. This includes detailing the algorithms, API settings, and decision-making criteria employed during data scraping. Such transparency not only enhances the reproducibility of research but also allows for a critical evaluation of the methodologies used, promoting trust and verifiability in the findings. This is not limited to only including timestamps for the collecting periods but also the selected settings/features/attributes of the APIs and other relevant scraper features used. Typically, there is a routine expectation for transparency regarding the process of subjective data collection, especially in human-based methods. However, this level of scrutiny is often overlooked when it comes to automated methods.

On the other hand, this responsibility could be shifted or partially shared if the data are not collected by the authors themselves but are provided by an external entity such as a company specialized in media analysis and scraping or even the news outlet itself. In the latter case, one then has to trust the outlet that they provide all the news stories on the topic they deem relevant. Additionally, in both the former and latter cases, the data collection becomes a true black box as reproducing the data collection is not possible based on solely the research article.

While our study concentrates on the frequency of climate change articles, we acknowledge that this is a mere slice of the narrative. The visibility and prominence given to these articles — such as front-page placement or feature positions on websites — play a crucial role in shaping public discourse. Future research could enrich our understanding by incorporating these dimensions, potentially utilizing sophisticated tools to analyze digital replicas and virtual formats for a more holistic picture of media influence.

Finally, we highlight the importance of securing public non-commercial databases collecting and storing media data. As media conglomerates and social media companies apply stricter commercially based data policies, such public databases become increasingly important both for manual and automated approaches.

References

Tanja Aitamurto and Seth C. Lewis. 2013. Open innovation in digital journalism: Examining the impact of Open APIs at four news organizations. New Media & Society, 15(2):314–331.

Ralf Barkemeyer, Frank Figge, Andreas Hoepner, Diane Holt, Johannes Marcelus Kraak, and Pei-Shan Yu. 2017. Media coverage of climate change: An international comparison. Environment and Planning C: Politics and Space, 35(6):1029–1054.

Annie Blatchford. 2020. Searching for online news content: the challenges and decisions. Communication Research and Practice, 6(2):143–156.

Max Boykoff, Meaghan Daly, Rogelio Fernandez Reyes, Jari Lyytimäki, Lucy McAllister, Marisa McNatt, Erkki Mervaala, Ami Nacu-Schmidt, David Oonk, and Olivia Pearman. 2019. World Newspaper Coverage of Climate Change or Global Warming, 2004-2023. Media and Climate Change Observatory Data Sets. Cooperative Institute for Research in Environmental Sciences, University of Colorado.

Maxwell T. Boykoff. 2011. Who Speaks for the Climate?: Making Sense of Media Reporting on Climate Change. Cambridge University Press, 1st ed.

Marcel Broersma and Frank Harbers. 2018. Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency. Digital Journalism, 6(9):1150–1164.

Axel Bruns. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Information, Communication & Society, 22(11):1544–1566.

Frederik De Grove, Kristof Boghe, and Lieven De Marez. 2020. (What) Can Journalism Studies Learn from Supervised Machine Learning? Journalism Studies, 21(7):912–927.

David Deacon. 2007. Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis. European Journal of Communication, 22(1):5–25.

Stacy Gilbert and Alexander Watkins. 2020. A comparison of news databases' coverage of digital-native news. Newspaper Research Journal, 41(3):317–332.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. Text as data: a new framework for machine learning and the social sciences. Princeton University Press, Princeton Oxford. ISBN: 9780691207544

Nick Hagar and Nicholas Diakopoulos. 2019. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. Media and Communication, 7(1):117–127.

Lindsay H. Hoffman. 2006. Is Internet Content Different after All? A Content Analysis of Mobilizing Information in Online and Print Newspapers. Journalism & Mass Communication Quarterly, 83(1):58–76.

Moaiad Khder. 2021. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and its Applications, 13(3):145–168.

Ville Kumpu. 2016. On making a big deal. Consensus and disagreement in the newspaper coverage of UN climate summits. Critical Discourse Studies, 13(2):143–157.

Claire Lauer, Eva Brumberger, and Aaron Beveridge. 2018. Hand Collecting and Coding Versus Data-Driven Methods in Technical and Professional Communication Research. IEEE Transactions on Professional Communication, 61(4):389–408.

Sumit K. Lodhia. 2010. Research methods for analysing World Wide Web sustainability communication. Social and Environmental Accountability Journal, 30(1):26–36.

Jari Lyytimäki. 2011. Mainstreaming climate policy: the role of media coverage in Finland. Mitigation and Adaptation Strategies for Global Change, 16(6):649–661.

Jari Lyytimäki. 2020. Environmental journalism in the Nordic countries. In In David B. Sachsman, & JoAnn Myer Valenti (Eds.) Routledge handbook of environmental journalism, pages 221–233. Routledge, London and New York. ISBN: 9781032336442

Dani Madrid-Morales. 2020. Using Computational Text Analysis Tools to Study African Online News Content. African Journalism Studies, 41(4):68–82.

Merja Mahrt and Michael Scharkow. 2013. The Value of Big Data in Digital Media Research. Journal of Broadcasting & Electronic Media, 57(1):20–33.

Eetu Mäkelä and Pihla Toivanen. 2021. Finnish Media Scrapers. Journal of Open Source Software, 6(68):3504.

Donica Mensing and Jennifer D. Greer. 2013. Above the Fold: A Comparison of the Lead Stories in Print and Online Newspapers. In Internet Newspapers, pages 283–302. Routledge, 0 ed.

Murray State University, Vlad Krotov, Leigh Johnson, Murray State University, Leiser Silva, and University of Houston. 2020. Legality and Ethics of Web Scraping. Communications of the Association for Information Systems, 47:539–563.

Sanoma. 2012. Helsingin Sanomat to go to the tabloid format. Sanoma.com.

Andreas Schmidt, Ana Ivanova, and Mike S. Schäfer. 2013. Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. Global Environmental Change, 23(5):1233–1248.

Elisa Shearer and Amy Mitchell. 2021. News Use Across Social Media Platforms in 2020. Pew Research Center.

Lotta Sillanmäki. 2023. HS:n päätoimittaja vastaa kritiikkiin verkon ja printin erilaisista otsikoista: ”Ei mennyt ihan putkeen”. Yle.fi.

S.C.M. de S Sirisuriya. 2015. Comparative Study on Web Scraping. Proceedings of 8th International Research Conference, KDU.

Nicholas Stern, editors. 2007. The economics of climate change: the Stern review. Cambridge University Press, Cambridge, UK ; New York.

Pertti Suhonen. 1994. Mediat, me ja ympäristö. Hanki ja jää, Helsinki. ISBN: 9518916446

Tuula Teräväinen, Markku Lehtonen, and Mari Martiskainen. 2011. Climate change, energy security, and risk—debating nuclear new build in Finland, France and the UK. Energy Policy, 39(6):3434–3442.

Tommaso Venturini and Richard Rogers. 2019. “API-Based Research” or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. Digital Journalism, 7(4):532–540.

Huub Wijfjes. 2017. Digital Humanities and Media History: A Challenge for Historical Newspaper Research. TMG Journal for Media History, 20(1):4.

Tuomas Ylä-Anttila, Juho Vesa, Veikko Eranti, Anna Kukkonen, Tomi Lehtimäki, Markku Lonkila, and Eeva Luhtakallio. 2018. Up with ecology, down with economy? The consolidation of the idea of climate change mitigation in the global public sphere. European Journal of Communication, 33(6):587–603.

Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. Ethics and Information Technology, 12(4):313–325.

i For example, there are several bot accounts on Twitter that highlight changes made to newspaper articles such as "Editing The Gray Lady" or @nyt_diff that reveals changes made on the main page of the New York Times website.