On the Replicability of Data Collection Using Online News Databases

Mikaela Karstens, *The Pennsylvania State University – The Behrend College, USA*
Michael J. Soules, *Naval Postgraduate School, USA*
Nick Dietrich, *Ohio Wesleyan University, USA*

ABSTRACT News databases, such as Factiva and Nexis Uni, are vital for the construction of many commonly used datasets of political events because they provide researchers with access to thousands of diverse news sources. This article raises several issues with news databases that pose a threat to the quality and replicability of data-collection efforts. We recommend best practices for using news databases to gather event data.

Event data are central to the empirical study of politics. Whether through large event datasets, small boutique data-collection efforts, or detailed case studies using news stories, political scientists often rely on news reports to form their foundation of analysis. The advent of news databases including Factiva and Nexis Uni provides researchers with an unprecedented ability to construct datasets of political events. The Militarized Interstate Dispute (MID) dataset (Palmer et al. 2022); the Uppsala Conflict Data Program (UCDP) (Gleditsch et al. 2002; Pettersson and Öberg 2020); the Armed Conflict Location and Event Dataset (ACLED) (Raleigh et al. 2010); and the Mass Mobilization in Autocracies Database (MMAD) (Weidmann and Rød 2019) are examples of popular datasets derived from news databases.[1] Data-collection projects provide a service to the discipline by lowering the barrier to entry for quantitative analysis.

Whereas these databases aid in the collection of data, they also introduce an opaque process through which researchers' queries are transformed into a list of relevant documents. Underneath this black box, news databases make frequent, undocumented changes to search algorithms and content that may render replication of previous searches impossible. Additionally, databases have implemented increasingly restrictive legal and financial barriers that bar scholars with limited resources from coding data or investigating the source material for existing datasets. The commodification and restriction of database access pose a threat to the discipline if they are not adequately understood and addressed.

Mikaela Karstens *is a postdoctoral teaching fellow at The Pennsylvania State University–The Behrend College. She can be reached at mikaelakarstens@gmail.com.*
Michael J. Soules *is Donald R. Beall Defence Fellow at the Naval Postgraduate School's Defence Analysis Department. He can be reached at michael.soules@nps.edu.*
Nick Dietrich *is assistant professor of data analytics at Ohio Wesleyan University. He can be reached at dietrich.nicholas@gmail.com.*

These issues are not new to everyone; those who specialize in the production of event data—including the European Network of Conflict Research and the Open Event Data Alliance—have discussed these issues for years.[2] However, there is a significant gap between those who are on the cutting edge of automated data production and those who use data to study substantive issues. We bridge this gap by identifying key issues and best practices for average users of event data or those constructing smaller event datasets.

WHAT ROLE DO NEW DATABASES PLAY IN THE CODING PROCESS?

Many event datasets use the same basic coding process: (1) gather a corpus of potentially relevant documents, (2) determine if those documents meet the criteria for inclusion in coding, (3) assign coding determinations following predetermined rules, and (4) review the results for accuracy.

News databases enable researchers to quickly locate a corpus of relevant documents by searching for particular sources and keywords. Figure 1 is a conceptual diagram of the coding process.

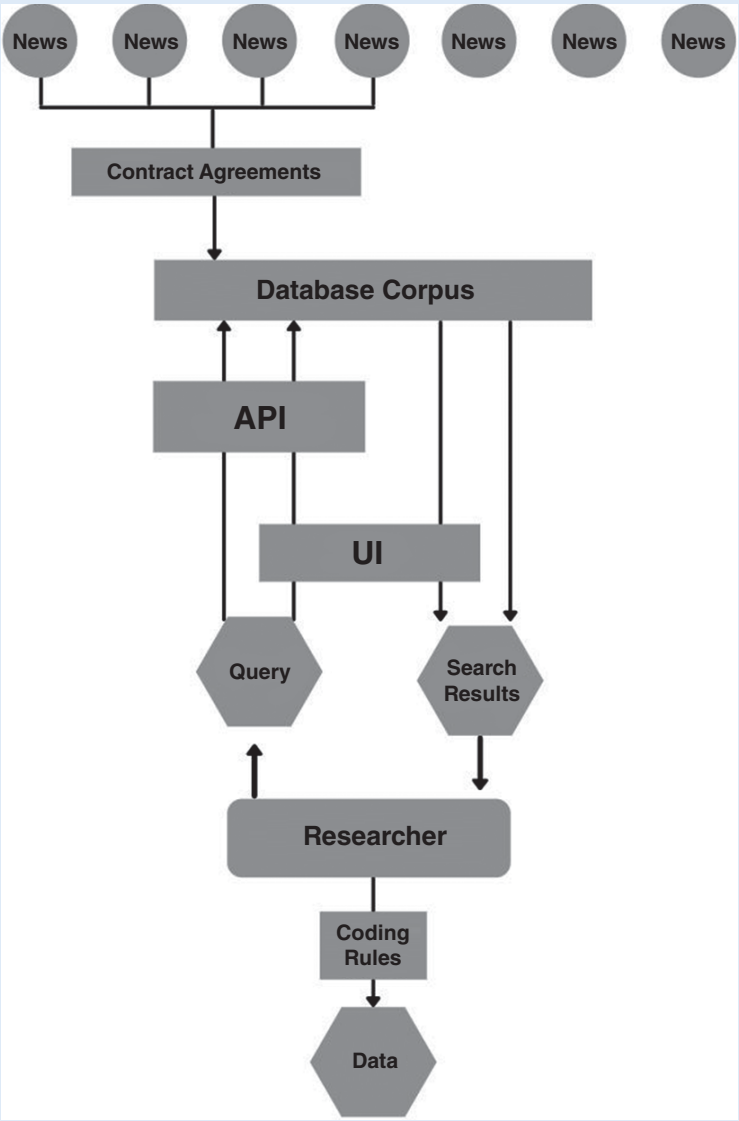
The application programming interface (API) allows the search application to communicate with digital archives. APIs can be used directly through applications written by a researcher or through command-line entries. Access to the API often is restricted to specialized, paid licenses. The user interface (UI) is the front-end interface often found on a database's proprietary programs. Information entered into the UI then is communicated through the API to the database corpus. These interfaces are managed by the databases themselves. For news databases, both the API and the UI serve as intermediaries between the search query and the articles that comprise the database.

The data-gathering process using news databases begins with the formation of the search query, which is defined as the combination of search terms, specified sources, and date ranges that a

doi:10.1017/S1049096522001317

Figure 1

Illustration of the Coding Process Using Newspaper Databases



researcher wants to examine. Although researchers can type simple search terms into database UIs, most are compatible with Boolean search terms. Boolean search is a logic-based approach to search queries in which keywords are strung together with operators including the words AND, OR, and NOT as well as brackets and quotations to form detailed queries.

After specifying the terms, a researcher enters queries into the database UI. The UI interprets the query and passes it along to the API, which—in turn—sends the request to the database. The database then returns a list of all documents that meet the search criteria to the UI, where the researcher can view the results. When using the API directly, the query can specify that results are returned directly into various file types. Once the researcher has access to the documents, they are filtered through to find which are relevant. In some projects, researchers begin reading through the documents, coding as they proceed. In other projects, the numerous results from their initial query require narrowing down through machine, manual, or hybrid approaches.

There is significant variability in how coding teams using news databases proceed, including how the corpus of documents is determined and whether analysis of selected articles is accomplished manually or through machine processing. Table 1 illustrates the process of a few major event datasets. Although coding projects vary in how they distill the corpus into data, they share a reliance on databases to reliably and accurately provide requested documents.

WHAT ARE NEWS DATABASES AND WHAT IS CHANGING?

The most prominent databases for researchers are LexisNexis products (e.g., Nexis Uni), Factiva, and Google News. In the case of proprietary databases such as Nexis Uni and Factiva, news agencies are contracted to allow inclusion of the creator's content

https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press

Table 1

Comparison of Selected Conflict Datasets

Dataset	News Database	News Coverage	Sources	Automated Classification
ACLED	Unspecified	Local, National, Global	Various	Unknown[11]
GTD	LexisNexis API	Local, National, Global	Various	Yes
MID4	Nexis Uni	Global	15	Yes
MMAD	LexisNexis Web Service Kit	Global	3 (AFP, AP, and BBC)	Yes
SCAD	Nexis Uni	Global	2 (AFP and AP)	No
UCDP	Factiva	Local, Global	Various	No

Sources: National Consortium for the Study of Terrorism and Responses to Terrorism (2018); Schrodt, Palmer, and Hatipoglu (2008)

within the database. When users search, they can view an article and, when permitted, download the complete text. These databases differ from open services such as Google News, in which the user is directed to the content creator's website. Access to proprietary databases often is granted through university libraries due to the substantial cost of obtaining access.[3]

There is a small but important body of literature that examines the databases themselves. This includes research that analyzes the similarities in news stories among databases (Weaver and Bimber 2008); the overlap in stories from the same news sources between databases and physical print (Ridout, Fowler, and Searles 2012); and the overlap in content availability between databases and microfilm (Youngblood, Bishop, and Worthington 2013). Existing work, however, has not examined how databases change over time or how those changes affect the discipline.

It takes a significant disruption to alert researchers that anything may be amiss. Most news-database changes are gradual and invisible to the end user and, as such, can go unnoticed. We identify the main changes to these services and their consequences. For clarity, we divide them into two categories: changes to terms of use and changes in functionality.

Changes to Terms of Use

Changes to terms of use are alterations in what is allowed under the database licensing agreements. They are driven largely by contracts between databases and content creators. The most relevant changes that we identified are article-download restrictions, article-retention limits, cross-institutional collaboration, and restrictions on machine- or crowd-assisted coding and processing. Table 2 summarizes these policies for the major news databases. Although some of these activities have always been prohibited, it is only recently that database managers have started policing violations.

Downloading source materials plays a central role in many coding efforts. Downloading articles allows coders to centrally store all source material and use alternative software to interact with the articles (e.g., Python). Although many databases allow downloading, functionality often is limited to one article at a time. This can be prohibitively slow if many articles are needed. When downloads are allowed, users may retain sources only for a finite amount of time, during which researchers must extract all needed information from texts before deleting them.[4] The ability to return to coding documentation to address questions regarding the data is crucial (Hensel and Mitchel 2015, 119). The inability to do so prevents replication and limits the ability to question or correct coding errors.

Current restrictions also prohibit sharing content with individuals outside of the user's institution.[5] Many event-data creators collaborate with researchers at other institutions. Additionally, this restriction prohibits datasets from sharing the text of their source documents if it is requested. Many coding projects use machine document processing to expedite the coding process. This is explicitly prohibited by most databases, and violations risk the academic license of the user's institution.[6] Users who want to use automated approaches must obtain alternative and often more-expensive licenses.

Changes in Functionality

Two changes to database functionality may cause researchers to miss events of interest: updated search-string interpretation and changing source lists. Search-string interpretation refers to situations in which the database UI interprets a researcher's query in unintended ways. We observed this happening through changing abbreviations for common sources, interpretation of dates, and requirements for inclusion/exclusion terms.

Most news database changes are gradual and invisible to the end user and, as such, can go unnoticed.

For example, the search string for the MID4 project abbreviated the source Agence France-Presse as "AFP." In coding the MID5 project, it was discovered that this shorthand no longer functions in Nexis Uni, resulting in "AFP" being excluded from the original batch of downloads for 2012. This exclusion was discovered in a subsequent audit of the search results. After correcting this and gathering all of the relevant AFP stories for 2012, 25 new incidents and three new disputes were discovered (Palmer et al. 2022). Had the MID team not performed multiple audits of its search results, these disputes would not have been included in the final dataset. The audit also revealed that the initial batch of downloads excluded stories from Interfax/ITAR-

https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press

Table 2

Comparison of Database Terms of Use for 2020

Dataset	Download	Retention Limit	Machine Processing	External Sharing
Factiva	Yes	30 Days	No	No
Google News	No	N/A	Yes	Yes
Nexis Uni	Yes	90 Days	No	No
NewsBank	Yes	Unspecified–Temporary	No	No
ProQuest	Yes	Unspecified	No	Yes–Limited

TASS for 2011. Eight new disputes were coded from those retrieved stories.[7]

Similarly, in Factiva, we encountered issues with dates in which specifying a range of dates (e.g., January 1–January 31) was interpreted in such a way that it excluded stories from the final day of the range (e.g., January 1, 12:01 a.m.–January 31, 12:01 a.m.). Updates such as these mean that researchers must perform periodic audits to ensure search strings are working as intended.

The sources available through the database also change over time. Contracts periodically expire or are renegotiated, resulting in sources dropping out of the database. This happens unannounced and intermittently. Additionally, the presence of the same source in two databases does not ensure access to the same publications. For example, Nexis Uni and Factiva both have Xinhua in their source lists but they contain different Xinhua publications.

Thus, search results may vary depending on factors hidden from the researcher. This is particularly disconcerting for projects that attempt to approximate the universe of cases. Whereas major events are unlikely to be omitted due to the substantial coverage they receive, smaller events that receive less media attention are more likely to be overlooked when a news story is lost in a database query.

To illustrate this issue, we tracked changes in search results on Factiva and Nexis Uni between February and October of 2019. Using search strings drawn from the MID Project (Palmer et al. 2022) as well as SCAD (Salehyan et al. 2012), we recorded the number of resulting articles for three periods: 2011–2018, January 2015, and January 2011 (figures 2 and 3). Both Factiva and Nexis Uni returned different numbers of articles for identical searches conducted over time within the same databases.[8] These week-to-week changes typically were small (i.e., <1% of total articles). We identified a small number of more significant changes, including one instance in which Factiva almost doubled the number of Associated Press stories returned in search results. The full results of this exploration are in the online appendix.

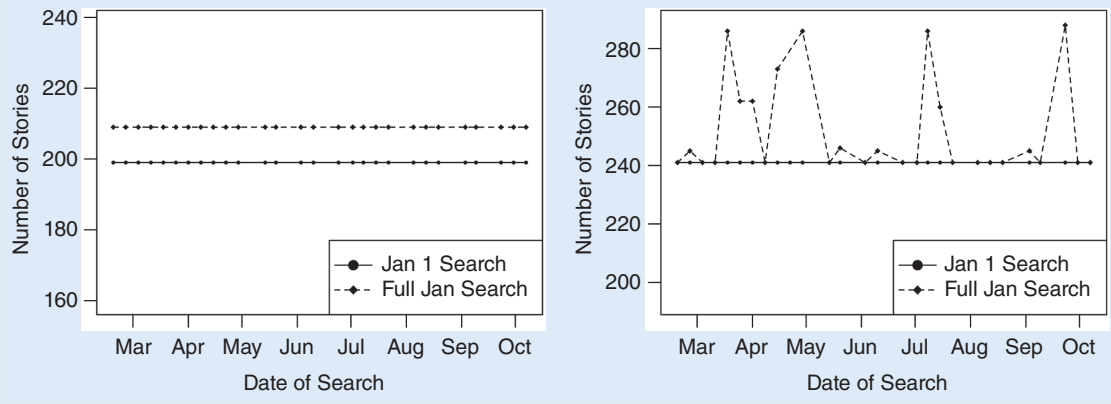
DOES DATABASE INSTABILITY AFFECT INFERENCE?

Instability in online news databases has the potential to affect datasets by preventing real events from appearing in the final dataset. An event that meets the coding criteria might be excluded if all source articles that document it are excluded from a database search. How often does this happen?

We are unable to provide a definitive answer without knowing which articles we have not observed. However, we can offer evidence about which events in published datasets are most at risk of exclusion and failure to replicate. The UCDP Georeferenced Event Dataset (GED) Version 21.1 (Sundberg and Melander 2013) records information about political-violence events. For events in the GED collected or updated since 2013, UCDP records the number of source articles used to code that event. Events with multiple sources should be more robust to fluctuations in database availability because the event still will be included, even if any single article is missing. Events with a single source, however, are more susceptible to these fluctuations because the omission of a single article could lead to the exclusion of that event. In the UCDP GED dataset, there are 157,303

Figure 2

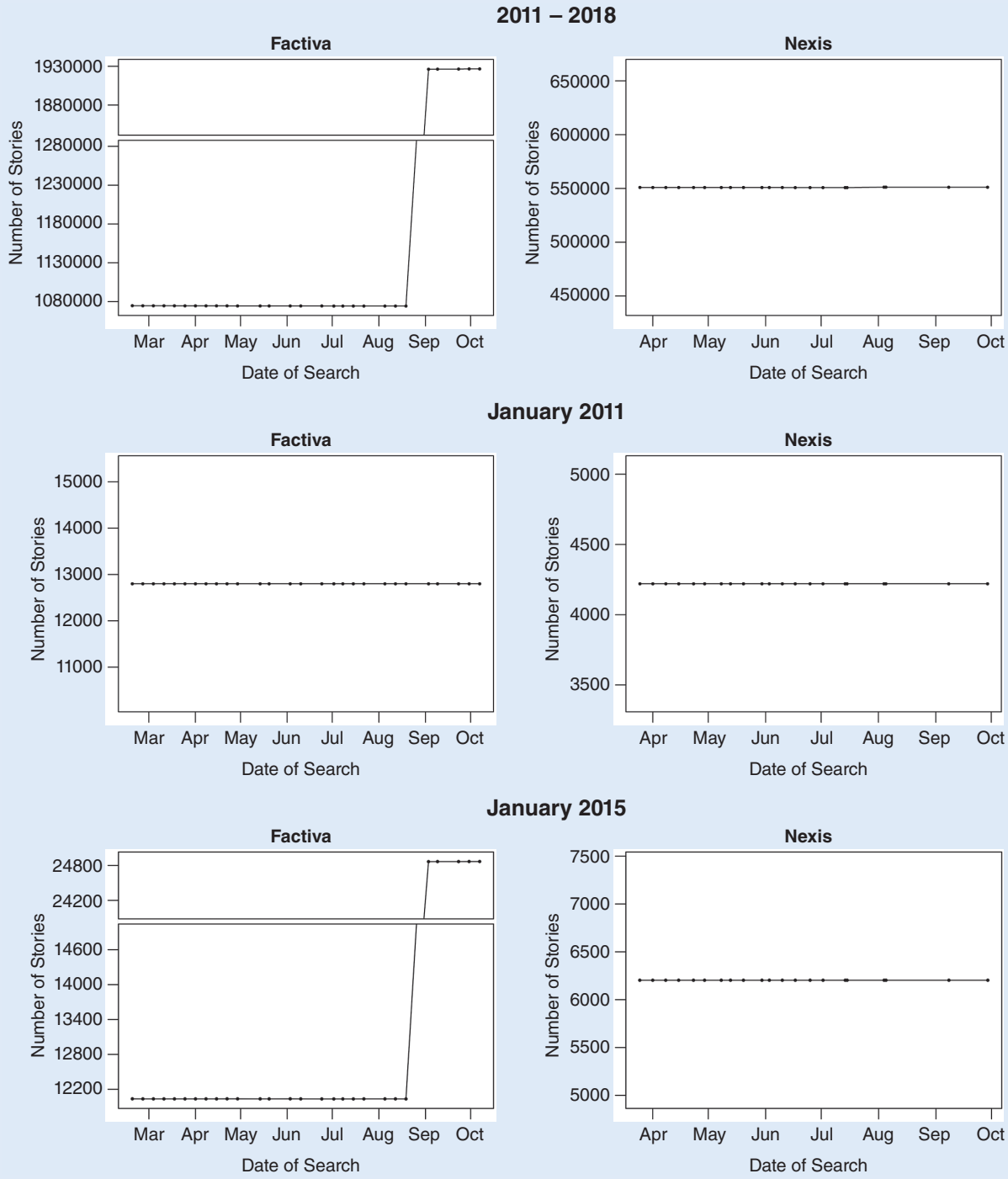
Variations in the Number of Factiva Search Results over Time



https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press

Figure 3

Variations in the Number of Associated Press Stories Retrieved Using the MID Search String



events with source information; 114,978 (73.1%) are coded from only a single source.

A second consideration is whether the inclusion or exclusion of events from a dataset affects *inferences* derived from it. We might conceptualize the exclusion of events due to database instability as stochastic measurement error unlikely to affect substantive conclusions of an analysis. However, there are reasons to believe that such errors are consequential for inference.

First, researchers often study a subset of events within a particular dataset. The omission of a single event from an entire dataset is unlikely to change the substantive conclusions gleaned. If, however, a researcher is performing an analysis of only a single country, then the omission of even a single event can be highly consequential. Such omissions *always* will have the effect of reducing the power of the analysis and increasing uncertainty around estimates.

https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press

Second, some events are more likely than others to be omitted because of database instability, which results in systematic bias. Events with fewer sources are at greater risk of omission from a dataset because there is a greater possibility that all source articles reporting on the event will be omitted from a news database. The number of sources reporting on an event is correlated with politically relevant characteristics such as the country in which the event occurred, the time at which it occurred, and whether it occurred in a rural or urban area. Dietrich and Eck (2020), for example, found that events occurring in Africa are less likely to be reported in media sources.

Third, individual events often are historically and politically significant on their own. Consider, for example, a researcher studying the history of territorial disputes between two countries. The omission of a border-skirmish event might make a year with one border skirmish appear to be a year of peace between the two countries. This is a difference of great narrative significance, despite it being an insufficient sample for statistical analysis.

...some events are more likely than others to be omitted because of database instability, which results in systematic bias.

IS API ACCESS THE SOLUTION?

API access and recent innovations in generating event data could alleviate some of these issues related to the terms of use and functionality of news databases. The MMAD (Weidmann and Rød 2019), which uses the LexisNexis Web Service Kit, is a good example of how APIs can be used effectively. MMAD contains data on political-protest events in autocratic countries. Data are generated at the level of an event report and then eventually aggregated into individual events. Aggregating event reports into events allows users to deal more effectively with issues related to inconsistent and missing information about the same event across multiple news reports (Weidmann and Rød 2019).

None of the other event datasets described in this article takes such an approach. We could argue that this approach might mitigate the issues discussed previously. For instance, MMAD uses a fairly coarse search string to download many articles and then uses computational methods to reduce the corpus of stories. Manual searches, in contrast, must use more complex search strings to produce a smaller corpus because manually downloading numerous stories is infeasible.

The greater complexity of manual search strings thus could be more vulnerable to the previously mentioned problem of the consequences of syntax choice changing without the user realizing it. Stated another way, the more terms that there are in a search string, the greater the opportunity for error. However, the use of APIs is not a perfect solution to this problem. Indeed, syntactical changes still could affect API users, necessitating periodic validation of search strings by researchers.

Another potential advantage of projects that use APIs (e.g., MMAD) is that sources are assigned numeric IDs, thereby avoiding the problems of changing sources names (e.g., "AFP" no longer working as an abbreviation for Agence France-Presse). Although this feature likely will help researchers with API access, there still are significant issues for those without such resources. Furthermore, API access does not solve the problem of source lists changing over time due to the expiration of contracts with specific news agencies. Thus, even those researchers with API access must remain vigilant about the sources from which their searches are pulling.

API access also can alleviate some of the issues associated with changing terms of use, particularly regarding the ability to download many sources. Again, however, API access often is prohibitively expensive, and the costs only continue to rise. Thus, researchers without such access may have trouble not only producing their event datasets but also replicating existing ones. Overall, whereas API access mitigates some of the issues discussed previously, it still (1) does not alleviate all of the discussed issues; and (2) only further widens the gap between the resources available to scholars at smaller and larger institutions to produce the types of research in which publishers often are interested.

BEST PRACTICES FOR CREATION

Some of these concerns cannot be addressed directly; however, simple best practices can help data gatherers to mitigate the impact of these changes. Although some of our recommended best practices are used by some existing datasets, there still are issues, in that (1) there is no evidence of all major event datasets implementing each of these practices; and (2) some datasets use a few of these suggestions but do not use others. Thus, we advocate for scholars to follow all of these best practices and to document their engagement in them, particularly given the increasing issues with using news databases.

1. *Select and Specify Your Sources.* Transparency regarding the source of information is critical in evaluating the veracity of data (Hensel and Mitchell 2015; Salehyan 2015.) Moreover, there are diminishing returns for including additional sources once the corpus reaches an acceptable level of coverage (Palmer et al. 2015).

Some event datasets (e.g., the MID, MMAD, and SCAD projects) already provide this information. However, other projects do not, making it difficult to determine how widespread this practice is. Moreover, even when datasets list the sources on which they drew, they do not always justify why these specific sources were selected. The MMAD project, however, provides a detailed justification for its source list.

Given the rising costs of news databases, it is becoming increasingly relevant for researchers to consider the value added from each additional source. As such, all data-gathering projects should provide users with documentation of which sources were used and when they were accessed. In cases in which researchers did not specify sources in the search string, they should document metadata regarding the sources that are used in their final coding.

2. *Validate Your Search-String Syntax.* It may take multiple iterations to land on the proper specification for a new search string. Researchers should review the results of their search after each adjustment and evaluate whether they are obtaining the expected results. For iterative data-collection projects, verification of the search string should be completed at the beginning of each iteration. Given that interfaces may change,

https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press

researchers must alter their search procedures to keep up with these changes. One test that researchers could perform is using their former search string to try to replicate a small part of a previous iteration of the dataset. To the best of our knowledge, validating search strings across different iterations of the same datasets is not a common practice—or, at the very least, there is no widespread documentation of such practices.

3. *Establish Explicit Procedures and Workflows.* Transparent data-collection procedures are crucial for the replicability of scientific research (e.g., Hensel and Mitchell 2015; King 1995). Some databases (e.g., SCAD) clearly explicate the specific steps used to generate the data[9]; other prominent event databases, however, do not. Coupled with the fluctuating nature of news databases, this makes replication difficult if not impossible.[10]

We suggest a list of relevant information that should be included in the description of data-generation processes. Other studies offer general best practices for replicability (e.g., Hensel and Mitchell 2015; King 1995; Salehyan 2015). Our advice focuses specifically on best practices regarding news databases. Although some existing datasets apply a few of these practices, we recommend that all event datasets should take each of the following steps and make such information publicly available:

- Specify which databases are used, if any.
- List the sources used or that no sources were specified.
- Provide justification for the sources chosen, including for the decision to search all sources instead of being selective.
- Include the exact search string(s) in replication materials.
- Include information on how articles are processed.
- State the dates of access.

4. *Conduct Periodic Checks of Your Results.* Researchers should periodically check their search results. For instance, in Nexis Uni, stories from Interfax News Agency are not available past December 31, 2010, even though a content listing provided by Nexis Uni, obtained in September of 2018, listed this source as being available from December 8, 1997, to the present.

Researchers should regularly check that all of their desired sources (1) appear in the data at all; and (2) appear every year from which the researcher(s) wants stories. Additionally, researchers should keep track of the number of results over time; the breadth and quality of coverage can fluctuate significantly (Rekatsinas, Dong, and Srivastava 2014).

Although fluctuations in the number of stories produced by a single news source are not inherently negative, drastic changes in availability from specific news sources could indicate consequential changes in the coverage of these sources. We are not aware of the practice of periodically checking search results to be widespread across major event datasets.

5. *Keep Up to Date on the Terms of Use.* Database terms of use regularly change in nontrivial ways. Researchers should keep up to date on whether their coding practices still align with database policy. In cases in which researchers need approval for automatic processing, batch downloading, or other prohibited techniques, they should investigate alternative licenses (e.g., API access). Given that many of the significant changes to the terms of use of news databases are relatively new, there has been little discussion of these issues in existing literature (for an exception, see Palmer et al. 2022).

CONCLUSIONS

News databases play a vital role in efforts to collect data on a variety of political events. Despite their central role in data collection, there has been little discussion of the problems with these resources and best practices for their use. Just as it is important for scholars to address potential biases in the sources they use, it also is important for researchers to directly confront issues with news databases to mitigate threats to replicability and validity.

This article discusses an array of issues related to these news databases. These changes pose an ethical risk to the field of quantitative political science. Given the restrictions to downloading and searching numerous sources, access to an API or batch-downloading services is necessary for many large data-collection efforts. The high cost of these services makes them inaccessible to scholars with limited resources or from less-privileged institutions.

This article is not intended to dissuade scholars from using news databases such as Nexis Uni and Factiva. Instead, the intent is to highlight these issues and discuss how they can be mitigated. Although news databases remain useful tools for building datasets, researchers must acknowledge and adapt to logistical problems when using them and the complications that they introduce to replicability.

ACKNOWLEDGMENTS

The authors are grateful to Chase Bloch, Brandon Bolte, Douglas Lemke, Cyanne Loyle, Roseanne McManus, and previous anonymous reviewers for their helpful comments on previous versions of this article. We also thank Andrew Dudash, Kayla Kahn, Michael Kenwick, Jeffrey Knapp, Glenn Palmer, and Kellan Ritter for their assistance in troubleshooting this issue as it was discovered and investigated. We also thank the Militarized Interstate Dispute Project for allowing us to share data and insights gathered through the process of coding the MID 5 dataset.

The high cost of these services makes them inaccessible to scholars with limited resources or from less-privileged institutions.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S1049096522001317.

CONFLICTS OF INTEREST

The authors declare that there are no ethical issues or conflicts of interest in this research. ■

NOTES

1. At the time of this writing, many of these databases had been cited more than 1,000 times.
2. See, for example, notes about source access in Althaus et al. (2020, 3).
3. These prices change depending on the number of concurrent users, size of institution, and level of access. The databases are not transparent regarding these price structures.
4. Users typically can retain metadata (e.g., headline, date, and source) indefinitely. However, recovering stories from the database using this information is time consuming.
5. ProQuest is the exception, allowing "(…) minimal, insubstantial amounts of materials retrieved from the service" to be shared with third-party colleagues (ProQuest n.d., Example A, section 1.e.b.).
6. This was confirmed in a June 2019 conversation with an executive director at Dow Jones/Factiva.
7. All eight new MIDs in this batch were cases of Russian bombers intercepted by NATO fighter jets. These interactions would not have been coded without these previously missed stories.
8. We conducted these tests on the website UI for both Factiva and Nexis Uni, primarily due to the prohibitive cost of obtaining API access. Therefore, we are unable to specify how much of the variation we observed in the number of search results is present in API searches. Variation caused by the way that UI commands are translated into search results should be absent from API searches. Variation caused by changes in the underlying database (e.g., sources disappearing from the database due to a license expiring) will be present in both UI and API searches.
9. See, for example, the SCAD codebook at www.strausscenter.org/scad.html.
10. Although it is unlikely that anyone would want to replicate an entire dataset, it is not uncommon for researchers to want to review information regarding specific events, particularly outliers.
11. The ACLED coding procedures do not mention any automated processing.

REFERENCES

Althaus, Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. 2020. "Cline Center Historical Phoenix Event Data Variable Descriptions." V1.3.0, May 4. University of Illinois Urbana–Champaign: Cline Center for Advanced Social Research. https://doi.org/10.13012/B2IDB-0647142_V3.

Dietrich, Nick, and Kristine Eck. 2020. "Known Unknowns: Media Bias in the Reporting of Political Violence." *International Interactions* 46 (6): 1043–60. https://doi.org/10.1080/03050629.2020.1814758.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946–2001: A New Dataset." *Journal of Peace Research* 39 (5): 615–37.

Hensel, Paul R., and Sara McLaughlin Mitchell. 2015. "Lessons from the Issue Correlates of War (ICOW) Project." *Journal of Peace Research* 52 (1): 116–19.

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28 (3): 444–52.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). 2018. "Global Terrorism Database [Data File]." www.start.umd.%20edu/gtd.

Palmer, Glenn, Vito d'Orazio, Michael Kenwick, and Matthew Lane. 2015. "The MID4 Dataset, 2002–2010: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 32 (2): 222–42.

Palmer, Glenn, Roseanne W. McManus, Vito D'Orazio, Michael R. Kenwick, Mikaela Karstens, Chase Bloch, Nick Dietrich, Kayla Kahn, Kellan Ritter, and Michael J. Soules. 2022. "The MID5 Dataset, 2011–2014: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 39 (4): 470–82.

Pettersson, Therese, and Magnus Öberg. 2020. "Organized Violence, 1989–2019." *Journal of Peace Research* 57 (4): 597–613.

ProQuest. "Terms and Conditions." Accessed February 12, 2021. https://about.proquest.com/about/terms-and-conditions.html.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47 (5): 651–60.

Rekatsinas, Theodoros, Xin Luna Dong, and Divesh Srivastava. 2014. "Characterizing and Selecting Fresh Data Sources." In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 919–30. Snowbird, Utah: June 22–27. https://doi.org/10.1145/2588555.2610504.

Ridout, Travis N., Erika Franklin Fowler, and Kathleen Searles. 2012. "Exploring the Validity of Electronic Newspaper Databases." *Exploring Journal of Social Research Methodology* 15 (6): 451–66.

Salehyan, Idean. 2015. "Best Practices in the Collection of Conflict Data." *Journal of Peace Research* 52 (1): 105–109.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38 (4): 503–11.

Schrodt, Philip A., Glenn Palmer, and Mehmet Emre Hatipoglu. 2008. "Automated Detection of Reports of Militarized Interstate Disputes Using SVM Document Classification Algorithm." Boston: Annual Meeting of the American Political Science Association.

Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–32.

Weaver, David A., and Bruce Bimber. 2008. "Finding News Stories: A Comparison of Searches Using LexisNexis and Google News." *Journalism/Mass Communication Quarterly* 85 (3): 515–30.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2019. "The Internet and Political Protest in Autocracies." Chapter 4. Oxford: Oxford University Press.

Youngblood, Norman E., Barbara A. Bishop, and Debra L. Worthington. 2013. "Database Search Results Can Differ from Newspaper Microfilm." *Newspaper Research Journal* 34 (1): 36–49.

https://doi.org/10.1017/S1049096522001317 Published online by Cambridge University Press