Achieving Transparency, Traceability, and Readability with Human-Coded Data

Amanda B. Edgell, *University of Alabama, USA*
Jean Lachapelle, *Université de Montréal, Canada*
Seraphine F. Maerz, *University of Melbourne, Australia*

ABSTRACT   Many important questions in political science require the use of human-coded data or information that has been systematically ordered and quantified by a human being from qualitative sources. This article discusses challenges and recent innovations in collecting and documenting human-coded data. We review five datasets produced within the past 10 years and also reflect on our experiences in collecting a quarterly dataset that tracked state responses to the COVID-19 pandemic. We argue that scholars can deliberately produce and publish theoretically grounded human-coded data in an accessible format that promotes transparency, traceability, and readability. We highlight several ways that scholars are already doing this, including narratives, source lists, and coding justifications that enhance the quality of their human-coded datasets. We also discuss common issues during coding and how technological innovation through interactive web-based platforms can improve the documentation of coding decisions.

Despite recent advances with automated web-scraping (Ulbricht 2020) and computer-assisted text analysis (Lucas et al. 2015), many important questions in political science still require data that have been systematically ordered and quantified by a human being from qualitative sources.[1] This type of data collection often is time-consuming, with difficult decisions on the researcher's part. After all, seemingly objective facts (e.g., the number of journalists imprisoned by a government) require subtle judgment from coders who must evaluate the credibility of claims made by different sources and clearly define the concept. For this reason, human-coded data face valid criticism for being expensive and difficult to reproduce (Benoit et al. 2016; Minhas, Ulfelder, and Ward 2015).

For example, a recent special issue of *PS: Political Science & Politics* on democratic backsliding reflects on the challenges of producing valid measures of social science concepts using human-coded data. Little and Meng (2024) argue that the current perception of global democratic backsliding could reflect coder biases rather than real regime change. They support their claims with an empirical illustration that relies on more objective data. However, Knutsen et al. (2024) argue that such seemingly objective indicators typically also rely on human decisions, are equally prone to human bias, and that the measurement of complex latent concepts like democracy requires substantial human judgment.

This article's main contribution is to provide a set of guiding principles for collecting human-coded data to mitigate problems that arise from subjectivity in coding. We emphasize transparency, traceability, and readability as three goals that researchers should

Corresponding author: Amanda B. Edgell (iD) *is assistant professor of political science at the University of Alabama and a research associate with the Varieties of Democracy (V-Dem) Institute. She can be reached at abedgell@ua.edu.*

Jean Lachapelle (iD) *is assistant professor of political science at the Université de Montréal and a research associate with the Varieties of Democracy (V-Dem) Institute. He can be reached at jean.lachapelle.1@umontreal.ca.*

Seraphine F. Maerz (iD) *is lecturer of political science at the University of Melbourne and a research associate with the Varieties of Democracy (V-Dem) Institute. She can be reached at seraphine.maerz@unimelb.edu.au.*

aim for when producing human-coded data. Whereas transparency has received increased attention from political scientists during the past 20 years, traceability and readability have not been theorized, to our knowledge. Yet we believe that they are equally important.

To illustrate our main points, we review five datasets produced within the past 10 years that have made important contributions to their respective areas of inquiry. We analyze source documentation and justifications for coding decisions using online repositories. These datasets cover important research areas including democratization (Treisman 2020), revolutions (Beissinger 2022), regime types (Geddes, Wright, and Frantz 2014), coups (Chin, Carter, and Wright 2021), and resistance campaigns (Chenoweth, Pinckney, and Lewis 2018). We also reflect on our experiences of working on the Pandemic Backsliding project—a quarterly dataset produced during the COVID-19 pandemic—from which originated many of our thoughts about how to improve the quality of human-coded data (Edgell et al. 2021; see also Lachapelle et al. 2020).

CASE SELECTION

We intentionally sampled five large-N human-coded datasets created within the past 10 years. Our goal was not to provide a representative sample of all human-coded datasets but rather to learn from recent advances by leading scholars on diverse concepts important to the discipline. With this goal in mind, we sought to cover research on democracy, authoritarianism, and collective action, and we looked for datasets with extensive documentation accessible through online archives. As a result, we selected the following datasets:

- *Autocratic Regimes Dataset* (Geddes, Wright, and Frantz 2014): Provides coverage of political regimes from 1946 to 2010 using a categorical typology including democracy, monarchy, and personalist, party-based, and military regimes (or mixed types); identifies how each regime ends, if violence occurred during transition; and whether the succeeding regime also was autocratic.
- *Colpus, the Varieties of Coups D'état Dataset* (Chin, Carter, and Wright 2021): Codes military and nonmilitary coup attempts since 1946, differentiating between coups that significantly alter regime coalitions (i.e., regime-change coups) and those that preserve existing coalitions (i.e., leader-shuffling coups); also provides information about the targets and perpetrators of the coup attempts.

violent and nonviolent actors seeking to introduce political change; provides information on whether the movement was successful in achieving its goals.
- *Revolutionary Episodes* (Beissinger 2022): Covers 345 revolutionary episodes from 1900 to 2014 and provides information on the timing, goals, size, and forms of contention, as well as regime features, deaths, and outcomes.

We also reflect on our experiences as principal investigators for the Pandemic Backsliding project, a human-coded dataset measuring violations of democratic standards during COVID-19 (Edgell et al. 2021). The Pandemic Backsliding project differs from the other datasets reviewed herein because it aimed to measure events in real time (rather than historical cases). This introduced additional challenges, particularly given the nature of the pandemic, wherein basic facts often were challenged. Thus, this data-collection effort provides insight into how scholars can generate human-coded data during an evolving global emergency.

RECENT ADVANCES IN CODING HUMAN-CODED DATA

Human-coded datasets face validity and reliability challenges because they rely on judgments that often are difficult to document and reproduce. By validity, we mean that the indicator captures the concept (at least in part). By reliability, we mean that the indicator is measured the same way across cases and can be reproduced by another scholar based on the provided documentation (Adcock and Collier 2001). We argue that these challenges need not undermine the value of human-coded datasets. Careful attention to transparency, traceability, and readability can mitigate concerns about the validity and reliability of the coding process. The following sections discuss these goals for human-coded data and provide concrete examples to illustrate best practices.

TRANSPARENCY

Transparency refers to an openness about the rules and principles that govern data collection. As a starting point, transparency requires that researchers define the underlying concepts, including how they differ from related concepts. Transparency also requires careful documentation of the coding rules, typically accomplished in a publicly available codebook, which includes all instructions given to coders, the rules governing their interactions, and how the project addressed uncertainty. In addition,

Transparency refers to an openness about the rules and principles that govern data collection.

- *Democracy by Mistake* (Treisman 2020): Provides data on democratization episodes from 1800 to 2015 using more than 2,000 sources and congruence analysis to evaluate whether the democratization process was deliberate, unintended, or by mistake, with information on the specific mistakes and author's confidence in the sources and rating.
- *Nonviolent and Violent Campaigns and Outcomes (NAVCO) 3.0 Dataset* (Chenoweth, Pinckney, and Lewis 2018): Covers more than 100,000 hand-coded events of political dissent from 1991 to 2012, including the day-to-day methods and tactics used by

transparency pertains to the rules for identifying credible sources. How many sources are considered sufficient? What types of sources are considered credible? Overall, transparency requires "a full account of the procedures used to collect or generate the data" (Lupia and Elman 2014, 21), which is critical for establishing the validity and reliability of human-coded datasets.

Inter-rater reliability (IRR) and consensus are two common approaches to human-coded data. When coders do *not* communicate with one another, IRR and inter-coder reliability (ICR) tests often are considered the gold standard for achieving transparency

and testing the validity of the collected data *after* coders have submitted their coding. To achieve transparency with these approaches, researchers should provide documentation about the number of coders and the IRR/ICR test results, including how these results were calculated. Especially when coding sources to measure latent concepts (i.e., coding questionnaires for which responses require a substantial degree of interpretation), IRR and ICR are useful for identifying and leveling out subjective coding decisions. However, IRR and ICR tests are time- and labor-intensive. Although there have been first attempts to combine ICR with automated content analysis for more efficient cross-validations (Pennings 2011; Song et al. 2020), ICR (and IRR) remains a challenging approach when resources are limited and for close to real-time data collection.

Alternatively, some studies take a consensus approach: coders communicate *during* the coding process and test the validity of the chosen material and its measurement *before* the actual coding starts. Through intensive deliberations, smaller test runs, and a collective decision-making process on source selection and coding procedures, the collaborative approach seeks to produce valid and reliable data while minimizing coding iterations. This approach is especially relevant when resources and human capacity are substantially limited or for data-collection endeavors with an ambitious timeline—for example, if the data should reflect (close to) real-time observations. To achieve transparency, such projects should provide documentation on the extent and nature of collaboration, as well as how the project achieved consensus—particularly for more uncertain or ambiguous cases.

Our survey of recent datasets reveals that documentation of concepts, coding rules, and procedures is increasingly detailed and transparent. Definitions of important underlying concepts often appear in journal articles introducing the dataset and in the codebook or other data documentation. For example, the Autocratic Regimes dataset provides a list of definitions and illustrative examples within the codebook (Geddes, Wright, and Frantz 2014). The codebook of the Colpus dataset includes a detailed discussion of alternative coups d'état definitions and highlights why and how the Colpus definition differs (Chin, Carter, and Wright 2021).

information about their collaborative coding strategies in their codebook and peer-reviewed journal articles (Beissinger 2022; Chenoweth, Pinckney, and Lewis 2018). Treisman (2020) outlined the coding process in the journal article and its online appendix (with additional documentation in the codebook). This included manually coding each case followed by an IRR test on a random subset. The codebook for the Colpus project includes extensive documentation of the coding rules with decision trees to guide readers through the process (Chin, Wright, and Carter 2021, 8, 11). Colpus uses one coder per datapoint and also provides an innovative measure of coder uncertainty based on the length of the case description, with the assumption that ambiguous cases require lengthier descriptions. Drawing on these practices, our Pandemic Backsliding project codebook included the coder questionnaire, and the online appendix to the article includes a copy of the instructions provided to research assistants (Edgell et al. 2021).

Although we observed a great deal of transparency across the datasets in our review, we also noted that authors sometimes provided important details only within the corresponding journal article, which makes it difficult for others to understand the coding process when these articles are not published through open access agreements (Chenoweth, Pinckney, and Lewis 2018; Geddes, Wright, and Frantz 2014; Treisman 2020). The codebook for the Autocratic Regimes Dataset also provides comparatively less information about how the coding happened (Geddes, Wright, and Frantz 2014). Thus, we recommend that authors provide full documentation of the coding process within open-access files to optimize transparency.

TRACEABILITY

Traceability allows data users to retrace in detail how coding actually happened by providing full details about which sources were used specifically for each datapoint, including justifications for individual coding decisions. Traceability is especially important when working with rich qualitative source material and historical events with competing interpretations (Cyr and Goodman 2024). Traceability underpins the work of historians who make many small decisions when reading through substantial

Traceability allows data users to retrace in detail how coding actually happened by providing full details about which sources were used specifically for each datapoint, including justifications for individual coding decisions.

Likewise, the data description for the Revolutionary Episodes dataset includes a thorough discussion of its conceptualization and how this relates to other concepts in the literature, with special attention to excluded cases (Beissinger 2022). For the Pandemic Backsliding project, we provide an explanation of the underlying concepts in the codebook and discuss them in more detail in an open-access journal article introducing the dataset (Edgell et al. 2021). The NAVCO 3.0 Dataset takes a similar approach, with limited definitions in the codebook and more detailed discussions in the corresponding journal article (Chenoweth, Pinckney, and Lewis 2018).

Most of the datasets we reviewed also provided information about the coding process with varying degrees of specificity. The NAVCO 3.0 and Revolutionary Episodes projects provide specific

amounts of archival documents (Farge 2013). Their decision rules often are quite inductive and difficult to systematize beforehand. Instead, they thoroughly document the archives that they consulted and reference the source material behind each of their claims, including the exact archive collection, document, and page numbers. Interested readers can verify the source material for themselves and determine whether the researcher's interpretation is substantiated.

Currently, most published datasets are accompanied by a thorough list of books, journal articles, and other sources used during coding. That said, not all datasets provide page numbers and coding justifications. Some provide a lengthy list of references, but this list does not allow users to retrace coding decisions because the reference is an entire book or article. This creates

difficulties in understanding the decision process behind each coding decision.

Several of the datasets we reviewed, however, provided comprehensive documentation of their sources for each datapoint, including specific page numbers for print sources. For example, the Colpus dataset provides a justification for each source and page numbers (Chin, Carter, and Wright 2021). The Revolutionary Episodes dataset includes a spreadsheet with complete references to all print sources, their page numbers, and a list of hyperlinks used to code each episode (Beissinger 2022). The Democracy by Mistake dataset has synopses totaling more than 2,300 pages, including all sources—with direct quotations and page numbers —that were used to code each case (Treisman 2020). In the Pandemic Backsliding project, we used computational tools to facilitate access to sources (Edgell et al. 2021).

Datasets that rely on web-based sources face a unique traceability problem because URLs may have a limited lifespan. Owners decide to redesign their websites, domain and security certificates expire, and the content may be removed altogether. As a result, "link rot" can reduce the long-term traceability of a project. We learned this lesson the hard way when we were coding our Pandemic Backsliding project dataset. Because we did not archive web-based sources in real time, we had to hire a separate team of research assistants at the end of the project to create permanent URLs for 6,423 webpages using Perma.cc and the Wayback Machine.[2] This was costly and time-consuming. It also reduced the traceability of our data because webpages may have changed, and 52 (1%) of our links were unrecoverable by the time we began to archive them. Whenever possible, we recommend that researchers plan to generate archived web pages as part of the ongoing data-collection process.

To summarize, traceability allows users to retrace the coders' footsteps by returning to the original sources and justifying difficult coding decisions. Based on this, users can understand more about how coding proceeded and the reliability of the data at hand. Backtracking the coders' footsteps also can help to verify the validity of sources and coding—that is, if the selected material and its processing sufficiently capture the implied concepts.

READABILITY

An "abundance problem" may arise from implementing traceability (Kim 2022). Consider a research project that endeavors to collect information about M political units on N dimensions: the resulting dataset will contain $M \times N$ datapoints. The number of sources and links that appear in the supplementary material will increase rapidly with the number of units and variables that researchers aim to collect. Indeed, asking coders to source and justify every coding decision can result in enormous amounts of unstructured metadata. Traceability diminishes if users cannot efficiently access coding justifications and source material. Thus, readability is a third guiding principle for human-coded data that aims to alleviate the problem that arises from making coding decisions traceable. By readability, we mean a system for presenting information about coding decisions and sources that is highly accessible and efficient.

Readability involves two fundamental questions: (1) how to structure the metadata and coder justifications; and (2) what format to use when distributing this information. Should the reference material be organized around the units or the variables collected? Organizing the data at the unit level may make the document easier to read, whereas organizing at the variable level facilitates comparisons among units for each variable. Furthermore, should the metadata be shared using a document file such as a PDF or some other format? This has implications for accessibility—for example, whether users have access to required software and knowledge about how to use it—and efficiency in finding information to retrace the authors' footsteps.

Most datasets surveyed for this article organize metadata at the unit level and share this data through a static document (e.g., PDF, Word, CSV, or XLSX file). Because most scholars are comfortable using these files, they are broadly accessible. However, finding information can be challenging as the number of documented variables and countries increases. For example, the Colpus dataset includes more than 1,900 pages of case narratives arranged into region-specific PDF files (Chin, Carter, and Wright 2021).[3] Similarly, Treisman (2020) included a synopsis of each democratization episode in a PDF file that is more than 2,000 pages long. The Revolutionary Episodes data provide narratives spanning more than 400 pages of Word files, an XLSM file with hyperlinks and print sources consulted for each episode, and a folder containing more than 200 PDF copies of source materials. The NAVCO 3.0 Dataset provides a master and 26 country-specific XLSX files with source documentation, coder comments, and arbitration decisions for achieving consensus (Chenoweth, Pinckney, and Lewis 2018).

The abundance problem sometimes forces scholars to focus on a few key variables rather than to document all decisions. For instance, the Autocratic Regimes dataset provides a detailed explanation for when a regime begins and ends but does not include a justification for coding the regime type (i.e., personalist, party-led, or military). As mentioned previously, this is understandable because the complexity of implementing traceability increases rapidly with each additional variable collected.

Concerns about readability encouraged us to think creatively about structuring and documenting the Pandemic Backsliding project dataset and, ultimately, guided our decision to move away from the traditional methods. We found that GitHub provides a useful interface for structuring a substantial amount of information underlying human coding decisions. We aimed to make this interface accessible for non-GitHub users by structuring the doc-

By readability, we mean a system for presenting information about coding decisions and sources that is highly accessible and efficient.

umentation into two main directories or folders, as illustrated in Figure 1. Users can access all of the coding decisions and sources for a particular country by clicking on the "by_country" folder or for a particular variable by clicking on the "by_question" folder. Each folder contains Markdown files corresponding to each country (in the "by_country" folder) or question (in the "by_question" folder), which GitHub automatically renders in the user's browser.

Screenshot of the Folder Structure on the Pandemic Backsliding GitHub Page

📁 by_country	update to v6.1	3 days ago
📁 by_question	update to v6.1	3 days ago
📁 code	update to v6.1	3 days ago
📁 codebook	Updated codebook 6.1	14 hours ago
📁 datasets	update to v6.1	3 days ago
📁 older versions	update to v6.1	3 days ago
📄 README.md	update readme v6.1	3 days ago

This makes it easy to produce readable documents from the raw data. To summarize, we structured the Pandemic Backsliding project GitHub to allow users to navigate through the documentation and access documents organized by country or question—similar to the folder structure found on most operating systems—thereby improving accessibility.

Thus, whereas static document files tend to be the norm, data science tools and online platforms can be especially useful for making metadata (i.e., all sources and coding justifications) broadly accessible. Online platforms such as GitHub make it possible to provide more information about coding than what currently is the norm, and they can organize information in a nonlinear manner. Researchers can click on datapoints of interest, access sources, and read coding justifications. Readability also might improve with the emergence of new technologies. Perhaps future researchers will design AI agents that will answer questions about coding procedures and they can refer to the appropriate piece of metadata based on user requests.

CONCLUSION

Despite the rise of automation in data collection, many concepts of interest to political scientists require human coders to translate qualitative content into quantitative values. Collecting such data involves several challenges, especially with contested concepts that are difficult to operationalize. This article reviews recent efforts to address these challenges. We observed substantive efforts in providing users with (1) full documentation of the rules and principles guiding the coding process (i.e., transparency); (2) detailed information on coding decisions and sources actually consulted (i.e., traceability); and (3) improved techniques and new data science tools for making metadata broadly accessible (i.e., readability).

Moving forward, we emphasize two key lessons drawn from recent human-coding efforts. First, as scholars increasingly rely on web-based sources, they should plan to address the issue of link rot from the early stages of data collection. For example, we were able to recover almost 99% of our online sources for the Pandemic Backsliding project. However, archiving these web pages during the data-collection effort would have been less resource-intensive, prevented the loss of several sources, and ensured that we captured the sources as they appeared in real time.

Second, although comprehensive justifications of coding decisions increase the transparency of datasets, their "offline" documentation in static documents (e.g., PDF, Word, and Excel files) can easily become overwhelming for users. Alternatively, online repositories such as GitHub facilitate data readability by allowing scholars to share their data, coding decisions, and sources with only a few clicks in an easy-to-use interface that mirrors the folder structure of almost all operating systems. However, using a standard GitHub repository also may incorrectly signal to users that knowledge of GitHub is required, thereby limiting readability. Therefore, we recommend that scholars who use GitHub to share their documentation files transform their repository into a more standard website layout using GitHub Pages[4] or to build an interactive dashboard using tools such as ShinyApps. We hope our review is useful to researchers who are interested in human-coded data, including those who seek to study ongoing events under time constraints.

ACKNOWLEDGMENTS

CONFLICTS OF INTEREST

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

NOTES

1. Despite the ongoing AI revolution and the development of tools such as ChatGPT, which are trained to generate human-like text responses, the need for human-coded data likely will persist. For example, Marquez's (2023) pilot study on using Large Language Models to generate measures of democracy demonstrates that whereas some of the results are surprisingly similar to human-coded measures, there nevertheless are major issues to be resolved in terms of accuracy and validation of AI-generated measures before human-coded data becomes obsolete.

2. See https://perma.cc and https://web.archive.org.

3. The Archigos project pioneered the narrative approach for documenting human-coded data (Goemans, Gleditsch, and Chiozza 2009). It is now quite common among human-coded datasets. We thank one anonymous reviewer for reminding us of this fact.

4. See https://pages.github.com.

REFERENCES

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46. DOI:10.1017/S0003055401003100.

Beissinger, Mark R. 2022. *The Revolutionary City: Urbanization and the Global Transformation of Rebellion.* Princeton, NJ: Princeton University Press.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. DOI:10.1017/S0003055416000058.

Chenoweth, Erica, Jonathan Pinckney, and Orion Lewis. 2018. "Days of Rage: Introducing the NAVCO 3.0 Dataset." *Journal of Peace Research* 55 (4): 524–34. DOI:10.1177/0022343318759411.

Chin, John J., David B. Carter, and Joseph G. Wright. 2021. "The Varieties of Coups D'état: Introducing the Colpus Dataset." *International Studies Quarterly* 65 (4): 1040–51. DOI:10.1093/isq/sqab058.

Chin, John, Joseph Wright, and David Carter. 2021. *The Colpus Dataset Codebook.* https://drive.google.com/file/d/1u1mtt4BYWlf6OoToqHeooGLmucj8GNLj/view.

Cyr, Jennifer, and Sarah Wallace Goodman. 2024. "Introduction." In *Doing Good Qualitative Research*, ed. Jennifer Cyr and Sara Wallace Goodman, 1–12. New York: Oxford University Press.

Edgell, Amanda B., Jean Lachapelle, Anna Lührmann, and Seraphine F. Maerz. 2021. "Pandemic Backsliding: Violations of Democratic Standards During COVID-19." *Social Science & Medicine* 285:1–10. DOI:10.1016/j.socscimed.2021.114244.

Farge, Arlette. 2013. *The Allure of the Archives.* New Haven, CT: Yale University Press.

Geddes, Barbara, Joseph Wright, and Erica Frantz. 2014. "Autocratic Breakdown and Regime Transitions: A New Dataset." *Perspectives on Politics* 12 (2): 313–31. DOI: 10.1017/S1537592714000851.

Goemans, Henk E., Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46 (2): 269–83. DOI:10.1177/0022343308100719.

Kim, Diana S. 2022. "Taming Abundance: Doing Digital Archival Research (as Political Scientists)." *PS: Political Science & Politics* 55 (3): 530–38. DOI:10.1017/S104909652100192X.

Knutsen, Carl Henrik, Kyle L. Marquardt, Brigitte Seim, Michael Coppedge, Juraj Medzihorsky, Amanda B. Edgell, Daniel Pemstein, Jan Teorell, John Gerring, and Staffan I. Lindberg. 2024. "Conceptual and Measurement Issues in Assessing Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 162–77. DOI: 10.1017/S104909652300077X.

Lachapelle, Jean, Steven Levitsky, Lucan A. Way, and Adam Casey. 2020. "Social Revolution and Authoritarian Durability." *World Politics* 72 (4): 557–600. DOI:10.1017/S0043887120000106.

Little, Andrew T., and Anne Meng. 2024. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 149–61. DOI:10.1017/S104909652300063X.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77. DOI:10.1093/pan/mpu019.

Lupia, Arthur, and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency: Introduction." *PS: Political Science & Politics* 47 (1): 19–42. DOI:10.1017/S1049096513001716.

Marquez, Xavier. 2023. *Using Large Language Models to Generate Democracy Scores: An Experiment.* https://xmarquez.github.io/GPTDemocracyIndex/GPTDemocracyIndex.html#how-good-are-ai-justifications.

Minhas, Shahryar, Jay Ulfelder, and Michael D. Ward. 2015. "Mining Texts to Efficiently Generate Global Data on Political Regime Types." *Research & Politics* 2 (3): 1–8. DOI:10.1177/2053168015589217.

Pennings, Paul. 2011. "Assessing the 'Gold Standard' of Party Policy Placements: Is Computerized Replication Possible?" *Electoral Studies* 30 (3): 561–70. DOI:10.1016/j.electstud.2011.05.001.

Song, Hyunjin, Petro Tolochko, Jakob Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G. Boomgaarden. 2020. "In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis." *Political Communication* 37 (4): 550–72. DOI:10.1080/10584609.2020.1723752.

Treisman, Daniel. 2020. "Democracy by Mistake: How the Errors of Autocrats Trigger Transitions to Freer Government." *American Political Science Review* 114 (3): 792–810. DOI:10.1017/S0003055420000180.

Ulbricht, Lena. 2020. "Scraping the Demos: Digitalization, Web Scraping, and the Democratic Project." *Democratization* 27 (3): 426–42. DOI:10.1080/13510347.2020.1714595.