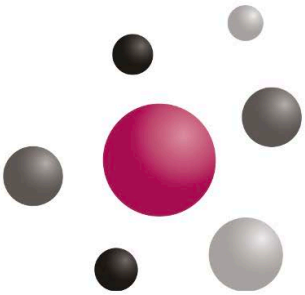

V-Dem INSTITUTE

Achieving Transparency, Reproducibility, and Readability with Hard-Coded Data: A Review

Amanda B. Edgell, Jean Lachapelle, and Seraphine F. Maerz

September 2023

Working Paper

SERIES 2023:145

THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (V–Dem) is a new approach to the conceptualization and measurement of democracy. It is co-hosted by the University of Gothenburg and University of Notre Dame. With a V–Dem Institute at University of Gothenburg that comprises almost ten staff members, and a project team across the world with four Principal Investigators, fifteen Project Managers, 30+ Regional Managers, 170 Country Coordinators, Research Assistants, and 2,500 Country Experts, the V–Dem project is one of the largest-ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

V–Dem Working Papers are available in electronic format at www.v-dem.net.

Achieving Transparency, Reproducibility, and Readability with Hard-Coded Data: A Review*

Amanda B. Edgell
Assistant Professor
University of Alabama

Jean Lachapelle
Assistant Professor
Université de Montréal

Seraphine F. Maerz
Research Fellow
Goethe University

*This research was supported by the Swedish Ministry of Foreign Affairs, grant number UD2020/08217/FMR and the University of Alabama Program for Middle East Studies. Seraphine F. Maerz received funding by the German Research Foundation, project number 421517935. We would like to thank Anna Lührmann, who was a Principal Investigator for the Pandemic Backsliding Project in 2020-2021; the core team of research assistants, which includes Sandra Grahn (manager), Ana Flavia Good God, Martin Lundstedt, Natalia Natsika, Palina Kolvani, and Shreeya Pillai; Abdalhadi Alijla, Tiago Fernandes, Staffan I. Lindberg, Hans Tung, Matthew Wilson and Nina Ilchenko, and the V-Dem country managers for their input during the pilot stage; Graham Baker, Cassidy Diamond, Roshan Malladi, and Christine Thompson, who generated the permalinks for this project; Waleed Hazbun, who provided funding for the permalinks coding; the University of Alabama Law Library, who hosted our Perma.cc account; Alexander Dukalskis and participants of the APSA and ECPR General Conferences in 2020, who provided valuable feedback. All errors and omissions are our own.

Abstract

Many important questions in political science require the use of "hard-coded" data or information that has been systematically ordered and quantified by a human being from qualitative sources. This working paper discusses challenges and recent innovations in collecting and documenting hard-coded data. We review five datasets produced within the last ten years and also reflect on our experiences working on the Pandemic Backsliding Project, a quarterly dataset tracking 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, reproducibility, and readability. We highlight several ways scholars are already doing this, such as through narratives, source lists, and coding justifications that enhance the quality of their hard-coded datasets. We also make suggestions for how technological innovation through interactive web-based platforms can improve the documentation of hard-coding decisions in the future.

Introduction

Despite recent advances with expert surveys (Coppedge et al. 2019), automated web-scraping (Ulbricht 2020), and computer-assisted text analysis (Lucas et al. 2015), many important questions in political science still require the use of hard-coded data. By "hard-coded", we mean information that has been systematically ordered and quantified by a human being from qualitative sources. The hard-coding process is far from straightforward; it is time-consuming and involves difficult decisions on the part of the researcher. For this reason, hard-coded data face valid criticisms for being expensive and difficult to reproduce (Benoit et al. 2016; Minhas, Ulfelder, and Ward 2015).

This working paper discusses several challenges and recent innovations in hard-coded data. We review five datasets produced within the last ten years that make important advances through narratives, documentation of sources, 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). These projects illustrate how scholars can deliberately produce and publish theoretically grounded human-coded data in an accessible format that promotes transparency, reproducibility, and readability. In addition, we reflect on our experiences working on the Pandemic Backsliding Project, a quarterly dataset produced during the COVID-19 pandemic (Edgell, Lachapelle, Lührmann, and Maerz 2021; Edgell, Lachapelle, Lührmann, Maerz, et al. 2021). We explain how we tackled the unique challenges of hard coding in quasi-real time during an evolving international crisis. We conclude with recommendations for future research based on our observations, particularly for those interested in hard-coding data from online sources.

Case Selection

For this working paper, we intentionally sampled five large-N hard-coded datasets created within the last ten years. Our goal is not to provide a representative sample of all hard-coded datasets but to learn from recent advances by leading scholars on diverse concepts important to the discipline. With this goal in mind, we sought to cover a range of research areas and looked for datasets with extensive documentation accessible through online archives. As a result, we selected the following datasets:

- **Autocratic Breakdown and Regime Transitions** (Geddes, Wright, and Frantz 2014): Provides coverage of political regimes from 1946 to 2010 using a categorical typology including democracy, monarchy, personal, party-based, and military regimes (or mixed types); identifies how each regime ends, if violence occurred during transition, and whether the succeeding regime was also autocratic.

- **Colpus, the Varieties of Coups D'état Dataset** (Chin, Carter, and Wright 2021): Codes military and non-military coup attempts since 1946, differentiating between coups that significantly alter regime coalitions (regime change coups) and those that preserve existing coalitions (leader shuffling coups); also provides information about the targets and perpetrators of the coup attempts.

- **The Nonviolent and Violent Campaigns and Outcomes (NAVCO) dataset** (Chenoweth, Pinckney, and Lewis 2018): Covers over 100,000 hand-coded events of political dissent from 1991 to 2012, including the day-to-day methods and tactics used by violent and non-violent 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, provides information on the timing, goals, size, and forms of contention, as well as, regime features, deaths, and outcome.

- **Democracy by Mistake** (Treisman 2020): Provides data on democratization episodes from 1800 to 2015 using over 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.

In addition to reviewing these five datasets, we also reflect on our experiences as PIs for the **Pandemic Backsliding** project, a hard-coded dataset measuring violations of democratic standards during COVID-19 (Edgell, Lachapelle, Lührmann, Maerz, et al. 2021; Edgell, Lachapelle, Lührmann, and Maerz 2021). The Pandemic Backsliding project differs from the other datasets reviewed here because it aimed to measure events in real-time (rather than historical cases), which introduces additional challenges, particularly given the nature of the pandemic, wherein some basic facts were often challenged. Furthermore, the data collection context required us to collaborate almost entirely online. Thus, this case study offers insights into how scholars can generate hard-coded data during an evolving global emergency.

Recent Advances in Coding Hard Data

Hard-coded data almost invariably involves a qualitative judgment on the researcher's part. Human beings must interpret qualitative texts – such as archival materials, government documents, and media reports – to operationalize big, important, and often multidimensional concepts into valid and reliable quantitative indicators. 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 documentation (Adcock and Collier 2001).

Because they rely on human judgments that are often difficult to document and reproduce, all hard-coded datasets face validity and reliability challenges. Seemingly objective measures, such as whether a protest movement employs violence, require some degree of subjective assessment on the part of the coders. Consider the 2011 Egyptian revolution. Although the event is typically described as a non-violent uprising, it involved many violent attacks on security forces, including looting, arson, and Molotov cocktails (Ketchley 2017, 30). Describing this event as non-violent requires making a subtle distinction between more spontaneous forms of collective violence that involve attacking security forces with makeshift weapons and other more organized forms of violence in which actors train in weapons and guerrilla tactics.

As this example illustrates, qualitative judgment in hard-coded data is inescapable for many concepts and research questions that interest political scientists. However, these challenges need not undermine the validity and reliability of hard-coded datasets. From our review of five recently hard-coded datasets, careful attention to transparency, traceability, and readability can help mitigate concerns about the hard-coding process.

Transparency

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 hard-coded datasets. To achieve transparency, many of the datasets we reviewed include a narrative about how the researchers interpreted the qualitative sources to arrive at each quantitative value observed in the data. For example, the autocratic regimes dataset (GWF) explains why an authoritarian regime begins and ends in a particular year based on historical references (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). The NAVCO project provides general explanations of concepts and specific information about its coding strategies in each section of its codebook (Chenoweth, Pinckney, and Lewis 2018). The Democracy by Mistake dataset gives rich contextual detail and assigns a confidence score to each rating (Treisman 2020). These approaches to transparency allow users to reconstruct the logic behind each coding decision.

Traceability

Traceability allows users to retrace the authors' footsteps by returning to the original sources. This requires more than just an overall list of references for the dataset; it requires details on the specific sources used for each data point. Increasingly, we find that

scholars provide thorough documentation of their sources for each data point, 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 comes with synopses totaling over 2300 pages, including all sources - with direct quotations and page numbers - used to code each case (Treisman 2020). These efforts at traceability enhance the reliability of hard-coded data by allowing other researchers to reproduce the coding process. When combined with transparency, traceability also helps users assess the validity of the data, especially for borderline cases, where a data point could reasonably be classified into different categories.

Readability

Providing a written justification and sources for every data point creates abundant information. Researchers must then decide how to organize and disseminate that information. Organization is key for validity and reliability, especially when coding in large teams. Coders often need to revisit their codings after criteria are clarified, and they need to ensure consistency with other coders. To achieve transparency and traceability, researchers must also decide how to disseminate information about their coding choices and sources to the wider public. We find that ensuring *readability* provides a solution to both the challenge of organization and dissemination in hard-coding projects. By readability, we mean a system for presenting information about coding decisions and sources that is highly accessible and efficient for both coders and users.

Most datasets use pdf or Excel files to summarize coding decisions. Some documents can easily reach several hundred pages long. Although the information is accessible using keyword searches, this format makes it impractical to justify all coding dimensions, especially when many variables are involved. Invariably, some data points are left out, as the task of providing a written justification for each data point can become unmanageable. Given these challenges, we see great potential in online interfaces and interactive tools that can efficiently present information about coding decisions and sources.

Hard Data in Hard Times: The Pandemic Backsliding Project as a Case Study

This section provides an in-depth case study of the Pandemic Backsliding project to illustrate how hard-coded data can achieve transparency, traceability, and readability even when collected during comparatively hard times (Edgell, Lachapelle, Lührmann,

and Maerz 2021). During the first year of the COVID-19 pandemic, we worked in a small research team to compile a quarterly dataset that tracks government responses to COVID-19. We were not alone in this effort; several other teams of researchers hard-coded COVID-19 policies around the world, such as lockdowns, school closures, and mask mandates (Cheng et al. 2020; Hale et al. 2021). However, the Pandemic Backsliding project aimed to assess whether states' responses to COVID-19 violated democratic norms. This required us to conceptualize democratic standards for emergency measures and to develop a method for collecting data across the world while the pandemic unfolded.

To do so, we worked with a team of trained research assistants who applied a common questionnaire to all countries. We developed this questionnaire based on a careful reading of international human rights law and theories about democracy during states of emergency (United Nations 2020; Lührmann and Rooney 2021). In April 2020, we ran a pilot study with crowd-sourced coders using our professional networks and the broader network of scholars involved with the Varieties of Democracy (V-Dem) Project.[1]

The pilot stage allowed us to fine-tune the questionnaire to reduce coder judgments and eliminate items outside the scope of our research question. Afterward, with funding from the Swedish Ministry of Foreign Affairs, we recruited and trained a team of research assistants who worked full-time on the project. The research assistants coded each round of the data drawing on reliable open-source online references, including government and non-governmental organizations' websites and media outlets. They documented all coding decisions with a brief narrative and provided links to their sources. We published this information through the project's Github repository (Edgell, Lachapelle, Lührmann, Maerz, et al. 2021). We also published quarterly policy briefs on the V-Dem website to provide regular updates to the public and scholarly community on our findings (Lührmann, Edgell, and Maerz 2020; Kolvani, Pillai, et al. 2020; Edgell et al. 2020; Kolvani et al. 2021; Kolvani, Lundstedt, Maerz, et al. 2020).

The Pandemic Backsliding project aimed at collecting hard data on a rapidly evolving political situation under heavy time constraints. During this quasi-real-time data collection process, particular challenges arose concerning the principles of transparency, traceability, and readability. First, the pandemic substantially altered our working environment. Our team worked remotely for the entire data collection process, through Zoom video calls, email, and Slack messaging. Because of these extraordinary circumstances, during the very early stages of the project, we learned that we needed to maintain highly detailed documentation of concept formulations, measurement strategies, and coding justifications. This improved internal communication and increased the transparency of the data generation process.

1. This included V-Dem project and country managers, affiliates at its regional centers, and scholars working directly at the V-Dem Institute in Gothenburg, Sweden. Because the names of country coders for the V-Dem Project are confidential, we did not crowd-source from the pool of country experts who complete the surveys for the annual V-Dem dataset.

Second, the information collected was almost entirely Internet-based, instead of relying on historiographical and library sources. This produced two challenges: For one, online sources may be unreliable, and distinguishing between trustworthy and untrustworthy sources requires additional effort on the part of coders. To assist coders, we developed a source guide for coders on how to identify reliable online sources. We also learned the value of assigning the same countries to each coder for each coding period so that they could develop personal knowledge of reliable sources for specific cases.

Toward the end of the project, we encountered a traceability problem unique to datasets that rely solely on web-based sources: link rot. The urls for many web-based sources have a limited lifespan because owners decide to redesign their websites, domain and security certificates expire, and sometimes the content is removed altogether. Regardless of the reason, link rot reduces the long-term traceability of a project because users can no longer access the original source material. To tackle this challenge, after the final round of coding, we worked with another team of research assistants to replace the original urls in the documentation with permanent versions for nearly all of the sources in the Pandemic Backsliding project. First, we used Perma.cc to generate permanent urls for 6,141 webpages.[2] Perma.cc is a powerful tool, but has some limitations. It cannot archive "gated" material that requires a subscription or log-in, and also often cannot create permanent links to social media posts on Facebook and Twitter. For 212 urls that were no longer available (aka "dead links") or that could not be archived using Perma.cc, we were able to access an archived version on the Wayback Machine.[3] In total, we produced a permanent archive of 6,353 (98.9%) of the original 6,423 urls used for the project. Only 52 links were unrecoverable. Another 18 links remained live at the time of publication but could not be archived with permanent links.

Table 1: Links from the Pandemic Backsliding Project that could not be recovered using Perma.cc.

	Wayback	Live	Dead	Total
404 error	132	0	28	160
Server error	48	0	14	62
Security certificate error	15	0	8	23
Facebook	0	15	1	16
Twitter	11	2	0	13
Gated	4	1	0	5
Routes to homepage	1	0	1	2
Article expired	1	0	0	1

As summarized in Table 1, the most common reason for link rot in our case was a "404 error", which usually occurs when the page has been (re)moved by the site owner. We encountered this problem 160 times, about 5% of the total links used in the dataset.

2. https://perma.cc
3. https://web.archive.org

Fortunately, the Wayback Machine had archives of 132 of these pages. Other common causes of link rot include server (n=62) and security certificate errors (n=23), which usually occur when the owner has not paid to maintain the website. In most of these cases, the Wayback Machine also had an archive. We also could not create permalinks using Perma.cc for 16 Facebook posts and 13 Twitter posts, all but one of which was still live. The Wayback Machine had archives for most of the tweets (n=11), but no archive for the Facebook pages. Dead links we could not recover on Wayback Machine affected 35 out of 144 countries coded in the Pandemic Backsliding Project. Namibia had the most deadlinks because the security certificate was expired for its Parliament website (n=8), and the government's COVID dashboard returned a server error (n=1). Guatemala also had five deadlinks due to 404 errors, three of which were government websites, and two belonged to a law firm.

Finally, we updated the dataset on a quarterly basis to follow a rapidly evolving situation. Thinking ahead about improving readability helped with this process. For example, we found that research assistants' needed efficient access to data from previous versions and other coders. We also sought to make all coding explanations and sources easily accessible to the public with a few clicks of a button. These concerns encouraged us to think creatively about structuring and documenting the dataset and, ultimately, guided our decision to move away from the traditional pdf file. Instead, we found that Github provides a useful interface for structuring a large amount of information underlying human coding decisions. We aimed to make this interface as accessible as possible for non-Git users by structuring the documentation into two main directories or folders as illustrated in Figure 1.

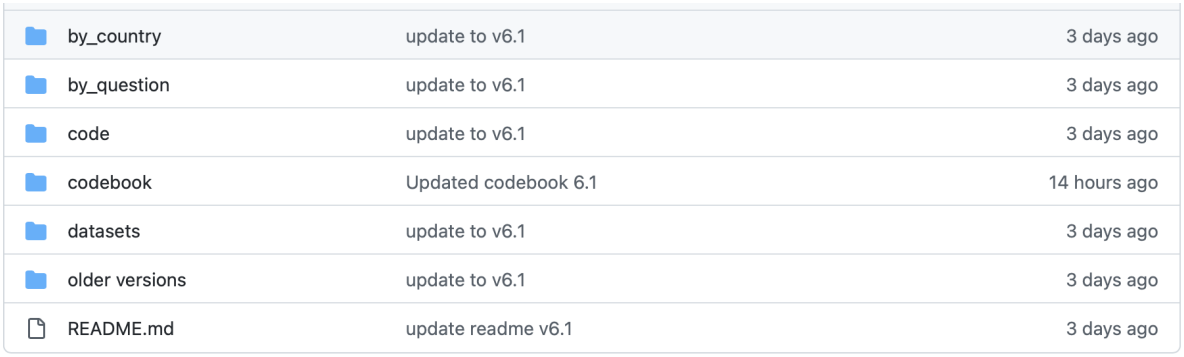

Figure 1: Screenshot of the folder structure on the Pandemic Backsliding Github page

Users can access all the coding decisions and sources for a particular country by clicking on the "by_country" folder or 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, making it easy to produce readable documents from the raw data. We wrote an R script to organize the data from the research assistants' spread-

sheets and create a readable Markdown document that is highly readable. In short, we structured the Pandemic Backsliding 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.

Conclusion

Despite the rise of automation in data collection, many concepts of interest to political scientists still require the collection of "hard" data, that is data that 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. In this working paper, we reviewed recent efforts to address these challenges. When looking at five hard-coded datasets released in the past decade, we observe substantive efforts to improve data transparency, traceability, and readability. We also reflected on our experience hard-coding the Pandemic Backsliding dataset, which presented novel challenges given that the data collection occurred in quasi-real-time during a global crisis.

Moving forward, we draw two valuable lessons from recent hard-coding efforts. First, as scholars increasingly rely upon web-based sources, they will need to tackle the issue of "link rot" from the early stages of data collection. For the Pandemic Backsliding project, for example, several of the urls for sources we used during the initial release of the dataset stopped working later on during data collection. This required time-intensive efforts to create permalinks for the original links that were still working. Although we were able to permalink or locate an archived version on Wayback Machine for nearly 99% of our online sources, archiving these webpages as we coded the dataset 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. This issue is especially important for projects that code ongoing events, and may be less relevant for projects that mostly use academic and historical sources, which are more easily traceable through print materials.

Second, while comprehensive justifications of coding decisions increase the transparency of datasets, their "offline" documentation in pdf files or Excel sheets can easily become overwhelming for users. As an alternative, online repositories like Github facilitate data transparency and accessibility by allowing scholars to share their data, coding decisions, and sources with just a few clicks in an easy-to-use interface that mirrors the folder structure found on almost all operating systems. However, using a standard Github repository may also incorrectly signal to users that knowledge of Git is required thereby limiting readability. Therefore, we recommend that scholars using Github to share their documentation files transform their repository into a more standard website

layout using Github Pages[4] or build an interactive dashboard to download using tools such as ShinyApps. We expect that this will increase user interaction and minimize the impression that knowledge of Git is necessary. We hope our review will be useful to future researchers interested in coding hard data, including those seeking to study ongoing events under time constraints.

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–546.

Beissinger, Mark R. 2022. *The Revolutionary City: Urbanization and the Global Transformation of Rebellion.* Princeton: 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): 278–295.

Cheng, Cindy, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. "COVID-19 government response event dataset (CoronaNet v. 1.0)." *Nature human behaviour* 4 (7): 756–768.

Chenoweth, Erica, Jonathan Pinckney, and Orion Lewis. 2018. "Days of rage: Introducing the NAVCO 3.0 dataset." *Journal of Peace Research* 55 (4): 524–534.

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): 1040–1051.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Joshua Krusell, Juraj Medzihorsky, Josefine Pernes, Svend-Erik Skaaning, Natalia Stepanova, Jan Teorell, Eitan Tzelgov, et al. 2019. "The methodology of 'Varieties of Democracy' (V-Dem)." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 143 (1): 107–133.

Edgell, Amanda B., Sandra Grahn, Jean Lachapelle, Anna Lührmann, and Seraphine F Maerz. 2020. "An update on pandemic backsliding: democracy four months after the beginning of the Covid-19 pandemic." *Policy Brief* 24:1–6.

4. https://pages.github.com

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.

Edgell, Amanda B., Jean Lachapelle, Anna Lührmann, Seraphine F. Maerz, Sandra Grahn, Palina Kolvani, Ana Flavia Good God, et al. 2021. *Pandemic Backsliding: Democracy During Covid-19 (PanDem), Version 6.1.* https://github.com/vdeminstitute/pandem.

Geddes, Barbara, Joseph Wright, and Erica Frantz. 2014. "Autocratic breakdown and regime transitions: A new data set." *Perspectives on politics* 12 (2): 313–331.

Hale, Thomas, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, et al. 2021. "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)." *Nature human behaviour* 5 (4): 529–538.

Ketchley, Neil. 2017. *Egypt in a Time of Revolution.* Cambridge University Press.

Kolvani, Palina, Martin Lundstedt, Amanda B. Edgell, and Jean Lachapelle. 2021. "Pandemic backsliding: A year of violations and advances in response to Covid-19." *Policy Brief* 32:1–6.

Kolvani, Palina, Martin Lundstedt, Seraphine F. Maerz, Anna Lührmann, Jean Lachapelle, Sandra Grahn, and Amanda B. Edgell. 2020. "Pandemic backsliding: Democracy and disinformation seven months into the Covid-19 pandemic." *V-Dem Policy Brief* 25:1–6.

Kolvani, Palina, Shreeya Pillai, Amanda B. Edgell, Sandra Grahn, Stefanie Kaiser, Jean Lachapelle, and Anna Lührmann. 2020. "Pandemic backsliding: Democracy nine months into the Covid-19 pandemic." *V-Dem Policy Brief* 26:1–6.

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): 254–277.

Lührmann, Anna, Amanda B. Edgell, and Seraphine F Maerz. 2020. "Pandemic backsliding: Does Covid-19 put democracy at risk." *V-Dem Policy Brief* 23:1–4.

Lührmann, Anna, and Bryan Rooney. 2021. "Autocratization by Decree: States of Emergency and Democratic Decline." *Comparative Politics* 53 (4): 617–649.

Lupia, Arthur, and Colin Elman. 2014. "Openness in political science: Data access and research transparency: Introduction." *PS: Political Science & Politics* 47 (1): 19–42.

Minhas, Shahryar, Jay Ulfelder, and Michael D Ward. 2015. "Mining texts to efficiently generate global data on political regime types." *Research & Politics* 2 (3).

Treisman, Daniel. 2020. "Democracy by mistake: How the errors of autocrats trigger transitions to freer government." *American Political Science Review* 114 (3): 792–810.

Ulbricht, Lena. 2020. "Scraping the demos. Digitalization, web scraping and the democratic project." *Democratization* 27 (3): 426–442.

United Nations. 2020. *International Covenant on Civil and Political Rights (ICCPR).* https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.