Digital Footprints and Data-Security Risks for Political Scientists

Colin Henry, *Vanderbilt University, USA*
Anita Gohdes, *Hertie School, Germany*
Cassy Dorff, *Vanderbilt University, USA*

ABSTRACT As greater shares of research and data are digitized, political scientists are increasingly confronted with questions pertaining to data security. Yet, data-management plans rarely evaluate the risks pertaining to a researcher's data across all project phases. This article highlights distinct risks related to key phases of the research process that deserve more attention by scholars. We emphasize risks during a project's inception and pre-data-collection phases, as well as risks associated with data publication and its afterlife. We discuss how shifts in political context and (re)newed politicization of topics can present new security risks for both the researcher and the researched communities long after a project has been completed. We provide a framework for recognizing and mitigating data risks, thereby contributing to the growing interest in data-security best practices.

In December 2019, the Indian Parliament passed the Citizenship Amendment Act (CAA), a law that provides the opportunity for undocumented migrants of a number of religious faiths to more easily obtain Indian citizenship (BBC 2019). The CAA was widely protested for the devastating consequences that it would have for the country's 170 million Muslims, who were excluded from the law (Kakkar 2020). India's ruling party also has been set on expanding the National Register of Citizens (NRC) to include all documented citizens of India. In the process, citizens have been excluded from the registry for failing to provide sufficient documentation. The NRC, in combination with the introduction of the discriminatory CAA, are examples of policy changes leading to a vastly increased politicization of information regarding ethnicity, religion, and migration history. Whereas data related to religion always has been sensitive, India's citizenship law shows that precautions taken around data collection—whether census-based, survey-based, or in the form of field notes—often cannot foresee future risks, even if they account well for current and past political dynamics.

Within political science research, risks associated with data collection and analysis are discussed mostly in the context of fieldwork, taking into account the human security of field workers, their research participants, and the security of their data and hardware (see, e.g., Cronin-Furman and Lake 2018; Gordon 2021; Grimm et al. 2020; Parkinson and Wood 2015; Wackenhut 2018). These debates are immensely important and inform safer and more ethical field research. However, they usually are restricted to studies that require Institutional Review Board (IRB) applications and collect some form of primary data.

This article contributes to important recent work that has highlighted data-security practices in the context of fieldwork (e.g., Miljanovic, Ververis, and Wissenbach 2020) by expanding on risks pertaining to *all* types of data collected across the research cycle. In particular, we focus on a project's inception phase and its afterlife. We argue that as a greater share of their work and data is digitized, political scientists across all subfields and methodological approaches must consider data security as a core component of their entire research process. The salience and sensitivity of information often changes over time, and data collected through processes that were deemed ethical and safe at one time suddenly could entail new threats (Knott 2019). Securing data used for research means attempting to exercise caution, care, and control over its use from inception of a research project until long after the results are published (Wackenhut 2018).

The following section discusses how the data-collection process produces a footprint that can be incorporated into threat evaluations. We then discuss the importance of anticipating the adverse effects the data may have not only on researchers and their research participants but also on future related target groups. The remainder of the article discusses practical ways to approach

Colin Henry *is a PhD candidate in political science at Vanderbilt University. He can be reached at colin.henry@vanderbilt.edu.*

Dr. Anita Gohdes *is professor of international and cyber security at the Hertie School in Berlin. She can be reached at gohdes@hertie-school.org.*

Dr. Cassy Dorff *is assistant professor of political science and affiliated faculty with the Data Science Institute at Vanderbilt University. Dr. Dorff can be reached at cassy.dorff@vanderbilt.edu.*

doi:10.1017/S1049096522000543

threats and risks to data security across these phases of the research cycle.

THINKING ABOUT DATA SECURITY BEYOND SPREADSHEETS AND PUBLICATION

This section expands the focus of data security beyond the most visible products of research—replication data and the final publication—to include threat modeling in all phases of scientific inquiry.

We argue that as a greater share of their work and data is digitized, political scientists across all subfields and methodological approaches must consider data security as a core component of their entire research process.

Understanding Data Footprints

Political scientists frequently conceptualize data in a traditional empirical sense wherein data represent a collection of observations, narratives, and records used for analysis. This definition admits a wide range of phenomena, such as square, "tidy," spreadsheet data, or "messy" databases gleaned from analogue and digital archives, to reams of interview notes written on actual paper. Whatever the format, data are subject-matter information at the heart of social inquiry. Some of these data contain personally identifiable information (PII), a subset of descriptive data that allows a sufficiently savvy user to isolate and identify a single individual. Examples include structured surveys, interview notes, and ethnographic observations. PII could be used to target identified individuals, but it frequently is used to connect other types of "data" to individuals. Both instances showcase the risk that data may pose to the privacy of the people from whom it is derived—either voluntarily or involuntarily. Scholars are duty-bound to protect human subjects, and vulnerable populations or individuals may be endangered by publication of politicized PII data.

However, these two common intuitions about data and their associated risks capture only a small portion of the total universe of data and their visible or implied hazards. Researchers themselves produce metadata throughout their research process. Every decision and action taken—including preliminary research design —produces a unit of information, or footprint, and it often is difficult to anticipate the threat method produced by each decision. Consider, for example, conflict researchers studying extremist groups' propaganda on social media. Although these data might not require researchers to leave their desk, numerous data-collection steps—including social media log-in credentials, application programming interface access tokens, and the identification of usernames—generate potentially vulnerable data-access points.

Anticipating Threats

Research questions and topics are as diverse as the data needed to answer them. As a consequence, anticipated threats related to researchers' processes differ widely. Borrowing from information security, *threat modeling* can establish a framework to evaluate risk in diverse research settings. Threat modeling denotes the process of identifying and evaluating potential threats, targets, and vulnerabilities in complex systems. Conventional use of threat modeling is widespread in computer science and information-technology environments but can be a useful tool for political scientists concerned about protecting sensitive data. With guidance from existing threat-modeling methodologies, the systematic evaluation of potential threat actors and the mitigation of potential attack methods can become as common a practice as budgeting for a grant proposal or building a survey instrument.

With guidance from existing threat-modeling methodologies, the systematic evaluation of potential threat actors and the mitigation of potential attack methods can become as common a practice as budgeting for a grant proposal or building a survey instrument.

Threat modeling can be categorized in three parts: (1) identifying assets, (2) rating threats, and (3) mitigating possible attack methods. Scholars can think of the research process and data life cycle as a system composed of a combination of software tools and research practices that should be subject to scrutiny by threat-modeling practices.

Research projects frequently focus on producing data through wholly new datasets, building novel combinations of existing datasets, or direct contact with research subjects during experiments and interviews. Often, anything produced this way can be considered an asset that should be protected through a stage of the research process. Although it is becoming more common for quantitative studies to provide raw data and replication code on publication, researchers may want to protect all of these data throughout the iterative scientific process. Research using interviews and other types of ethnographic data collection may be protected after collection and beyond publication—or even destroyed (Miljanovic, Ververis, and Wissenbach 2020). However, "assets" such as user account access to software-as-a-service (SaaS), email accounts, location information for researchers, and budgetary information also are candidates for threat modeling. Data collection is important, but information created at the beginning and the end of the research project should not be relegated to second-class status.

Many resources[1] are available to scholars unfamiliar with the process of securing digital information, and the ubiquity of computer-security challenges has led to a variety of reliable, open-source protection strategies. Solutions range from the simple (e.g., "do not reuse passwords across access points") to the complex and

https://doi.org/10.1017/S1049096522000543 Published online by Cambridge University Press

inconvenient (e.g., "air-gap all machines storing research data"). The following section describes a broader framework for researchers to begin evaluating the possible threats and risks that their data and data footprint might pose across core research phases.

EVALUATING THREATS AND RISKS ACROSS RESEARCH PHASES

To illustrate how researchers can better apply threat modeling to their research, we examine phases of the research cycle and highlight the ways in which individual researchers can identify assets and rate threats to mitigate possible attacks to data security. We identify four key phases: (1) project inception and pre-data collection, (2) data collection, (3) data processing and analysis, and (4) publication and data afterlife. Table 1 is a checklist to guide a researcher's risk evaluation throughout the research pipeline. To illustrate their importance, the following discussion focuses mainly on the first and fourth phases of research.

Table 1

Checklist to Guide Risk Evaluation Across Research Phases

Project Phase	Questions to Guide Risk Evaluation
(1) Inception and Pre-Data Collection	→ Do your actions raise your profile or increase awareness about those who might become research participants? → Are you… • engaging in pre-planning conversations, interviews, or meetings? • traveling to a location that you plan to return to for further research? → publicizing your research intentions through institutional press releases, social media, and open-source code contributions?
(2) Data Collection	→ Has your ethics protocol/IRB[2]… • approval focused only on data security of direct research participants or interviewees? • protocol considered the protection of data generated by team members, including research assistants and field contacts such as drivers and translators? • protocol considered the ways in which transit counts as an important "place" for data breaches? → Have you considered keeping a list of all third-party communication tools and privacy policies that you plan to use during this time?
(3) Data Processing and Analysis	→ Have you… • brainstormed the ways in which your analysis of data might pose threats to reidentification of individuals or communities? • considered how your analysis might generate new threats to the research team due to the new nature of the data? → determined if your research involves a third-party host such as Github, transcription services, or Dropbox?
(4) Publication and Data Afterlife	→ How might the public release of your data… • affect others, possibly unintentionally? • be used in ways that could endanger research participants, yourself, or your team? • lead to future reidentification? • be useful for adversarial actors (e.g., repressive governments) to learn about challenger(s)? → Could data embargos alleviate possible concerns? → Which parts of the project "should" be replicable and publicly available and which should not? • Can you delete parts of your data (including metadata) that are not relevant for the replication of your results?

Project Inception and Pre-Data-Collection Phases

In the pre-data-collection phase of a project, researchers engage in brainstorming, formation of questions, early scoping discussions, literature review, and a general assessment of their research project's viability. During this phase, researchers unintentionally may disclose that they are working on a sensitive topic. Sensitive topics include issues in which data collection could endanger the research participants, the researchers themselves, or their immediate network. Early exploratory conversations and brainstorming sessions tend to be unprotected and informal, and they may produce metadata from interactions such as email conversations and online communications. In this phase, researchers also should be mindful about whether the informal information they gather warrants the same protections and care that more formal data gathering traditionally is given.

In this pre-data-collection phase, substantive topic knowledge or the process of identifying cases, organizations, and individuals for research helps researchers to compile a preliminary group of actors who may pose a data-security threat. Researchers also should use this phase of case selection to consider the threat capabilities and motivations of potential adversaries. Conversely, mitigation tools and strategies that are appropriate in some cases may expose participants to greater risk in others. Training and best practices suggested by local and institutional security experts should be balanced by consultation with academics, journalists, activists, and others working within the case-specific context of interest. Desk research that combines or transforms datasets should warrant consultation with case experts similar to fieldwork. Researchers planning to use facial-recognition software, for instance, might consult with technology ethicists to understand how data production might create risk for previous *and* current research participants.

Although these considerations unfortunately may rule out some case studies, we encourage researchers to take data-security threats as seriously as threats to human integrity. We extend the same caution to research designs that expose the data of scholars, research partners, and communities of study to undue risks. Specific recommendations that emerge from risk evaluation in this phase vary by case, but researchers should ensure that project

https://doi.org/10.1017/S1049096522000543 Published online by Cambridge University Press

members receive the training and knowledge needed to recognize security threats and to avoid them.

In this pre-data-collection phase, researchers also can compile a list of all third-party communication platforms relevant to the project and investigate (or ask university support to investigate) whether apps have access to features such as a smartphone's microphone and camera. Communication platforms should not be vetted only for research projects that include a fieldwork component; project teams that work on sensitive or politicized topics that assemble digital and secondary data sources also should give serious consideration to the risks involved in their communication protocols.

Finally, in this phase, it also is beneficial to assess whether a project's data protection should include mapping key locations where researchers may be tempted to use public Wi-Fi or insecure networks or be exposed to other high-traffic locations (e.g., airports and public-transportation hubs).

Data-Collection Phase

After a researcher has moved into the data-collection phase, an iterative cycle between data collection and storage begins. Prior studies pertaining to fieldwork security provides important and detailed discussions of data-handling practices during the data-collection phase (see, e.g., Cronin-Furman and Lake 2018; Glasius et al. 2018; Grimm et al. 2020; Morgenbesser and Weiss 2018).

We emphasize that threats to researchers and subjects are not limited to situations in which data collection crosses country borders. We follow Grimm et al. (2020, 6), who referred to the field as "the environment where research takes place." Researchers' home countries may politicize or target work on vulnerable or marginalized communities, LGBTQ+ populations, racial minorities, and protest movements—to name only a few—or topics such as critical race theory, immigration, and infectious-disease prevention. This explicitly concerns researchers based in Europe and North America because illiberal democratic practices and policies are increasing in both regions. It is important to note that the sensitivity of topics may change over time. These topics may be broadly conceived of as "low-profile" or "acceptable" before starting the data-collection process, but their rapid politicization may provoke new dangers for both researchers and their research participants (Knott 2019). Data-security protocols therefore should consider security concerns beyond research topics traditionally viewed as high risk or sensitive.

How future actors use data is not always in the researcher's control.

As highlighted by prior work, anonymized interview transcripts nevertheless can reveal sensitive and damaging information about respondents (Cronin-Furman and Lake 2018; Glasius et al. 2018), even when researchers have an IRB-approved protocol for taking interview notes. This may be particularly true if transcripts were paired with travel itineraries and personal information.

Data Processing and Analysis

The data-processing and analysis phase of a project's life cycle involves hundreds of important decisions related to organizing data, data storage, and even data manipulation through anonymizing and merging records. In many cases, merging newly collected data with previously public information could generate an unintentionally sensitive asset associated with a higher threat potential for tracking individuals and reidentification. In this way, the data-processing phase can produce new data through data manipulation as well as new findings through analysis.

As new datasets are created, project leaders must be mindful of how new data assets and analytical results may be misinterpreted in good faith or misused in bad faith. The convention of applying disclaimers and limiting distribution of preliminary drafts of academic papers, reports, and books can be applied to datasets, visualizations, statistical tables, and qualitative results. By this phase in the research cycle and using theoretical or substantive case knowledge, scholars often will have identified subject organizations, communities, and individuals who may have a stake in newly created data assets and analyses. Moreover, we may believe that our processing and analytical work happens out of sight of these actors; however, conference proceedings, workshop materials, code repositories, data-storage facilities, and other public or semi-public sources often are only an Internet search away for a motivated adversary.

Researchers also should be cautious about engaging third-party vendors and using new computational infrastructure in this phase. Projects that successfully protected field researchers, study partners, and communities of interest may reverse these accomplishments by inadvertently expanding the data footprint in the analytical phase. For example, imagine a careful researcher who has gathered hours of recorded interviews from a sensitive group fearful of recrimination in the country in which the researcher lives. If the interviews are uploaded for transcription to a third-party software vendor without the researcher scrutinizing its data privacy policies, the interview subjects risk being exposed to identification and potential violence. Using SaaS usually entails sending data to a computer over which researchers do not have control. Reputable vendors should have sensible data-privacy policies, including end-to-end encryption, two-factor authentication, and limited or zero data retention. A researcher's institution may provide support for choosing a third-party vendor for data-analysis tasks that meets these standards.

Publication and Data Afterlife

Conventional protocol suggests that by the final phases of a research project, threats to data security have largely been identified and eliminated. This is not true, however, when we consider the ways that the meaning of data can change over time.

How future actors use data is not always in the researcher's control. Data collection is subject to scoping conditions, careful case selection, and a host of contextual decisions by researchers that frequently are explained in detail within the text of publications. As discussed previously, the meaning and salience of data can change over time as new topics and communities find themselves politicized. Researchers should consider what their data might be used for when stripped of this context. For

https://doi.org/10.1017/S1049096522000543 Published online by Cambridge University Press

example, data that previously were collected for non-research purposes or unrelated studies can be used in widely different research projects. Images scraped from Google and photographs clipped from YouTube videos recently became part of a US Government–funded database used to train for facial-recognition technology (Glaser 2019; Murgia 2019). Researchers may be able to mitigate this harm by integrating crucial contextual information within replication data through codebooks or disclaimers embedded in raw-data files. Another useful strategy is asking colleagues to write critical genealogies or histories of datasets that provide rich contextual information and can be circulated rapidly if a researcher's data becomes the focus of a misinformation campaign.

As scholars, we must think about data being stored properly, safely, and securely and with the option of deleting it in the future. The meaning of data can change quickly, and we must be aware of these changing dynamics. The political climate in a researcher's country of residence may change which type of research is permissible, what type of data may be collected, and with whom these data must be shared. For example, scholars who previously were free to pursue independent research have come under increasing pressure in countries such as Turkey and Hungary (Grimm and Saliba 2017). Researchers in the United States have found themselves on "watch-lists" detailing who allegedly "discriminate[d] against conservative students and advance[d] leftist propaganda in the classroom."

Publication outlets have their own guidelines about storage and maintenance of replication data. We encourage researchers to follow these guidelines when appropriate but also to reasonably question them. In some cases, researchers can request a "data embargo" so that data are to be released years after an agreed-on time. In addition, researchers can consider whether all of their data truly are required for replication. Should readers and analysts be able to reproduce the final dataset used in the analysis or simply use a final dataset to rerun key results? In many cases, researchers can post a streamlined, post-process dataset and provide other details on request. This would alleviate threats targeted at sensitive information as well as information that subsequently may aid in reidentification.

CONCLUSION

The demands of remote work and collaboration imposed by the COVID-19 pandemic highlight the need for digital-security literacy among political scientists. This article builds on prior studies of fieldwork risks and emphasizes aspects of data security that we believe deserve more attention from researchers. In particular, we argue that two phases of the research process that often receive less attention when considering data security—project inception and post-publication—can present a unique set of threats to researchers, participants, and communities. Finally, we emphasize that data insecurity can threaten desk research in different but important ways than fieldwork.

CONFLICTS OF INTEREST

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

NOTES

[1] For example, the Security Planner provides a personalized guide to better security planning that builds on expert advice. See https://securityplanner.consumerreports.org.

[2] The availability of IRBs varies greatly among countries. When IRBs are not available, researchers can rely on best practices, local ethical review boards, or create informal committees of scholars to review their ethics protocol.

REFERENCES

BBC. 2019. "Assam NRC: What Next for 1.9 Million 'Stateless' Indians?" *BBC News*, August 31. www.bbc.com/news/world-asia-india-49520593.

Cronin-Furman, Kate, and Milli Lake. 2018. "Ethics Abroad: Fieldwork in Fragile and Violent Contexts." *PS: Political Science & Politics* 51 (3): 607–14.

Glaser, April. 2019. "Palantir Said It Had Nothing to Do with ICE Deportations. New Documents Seem to Tell a Different Story." *Slate*, May 2. https://slate.com/technology/2019/05/documents-reveal-palantir-software-is-used-for-ice-deportations.html.

Glasius, Marlies, Meta De Lange, Jos Bartman, Emanuela Dalmasso, Aofei Lv, Adele Del Sordi, Marcus Michaelsen, and Kris Ruijgrok. 2018. *Research, Ethics, and Risk in the Authoritarian Field.* Berlin, Germany: Springer Nature.

Gordon, Eleanor. 2021. "The Researcher and the Researched: Navigating the Challenges of Research in Conflict-Affected Environments." *International Studies Review* 23 (1): 59–88.

Grimm, Jannis, Kevin Koehler, Ellen M. Lust, Ilyas Saliba, and Isabell Schierenbeck. 2020. *Safer Field Research in the Social Sciences: A Guide to Human and Digital Security in Hostile Environments.* London: SAGE Publications.

Grimm, Jannis, and Ilyas Saliba. 2017. "Free Research in Fearful Times: Conceptualizing an Index to Monitor Academic Freedom." *Interdisciplinary Political Studies* 3 (1): 41–75.

Kakkar, Jhalak M. 2020. "India's New Citizenship Law and Its Anti-Secular Implications." *Lawfare Blog*, January 16. www.lawfareblog.com/indias-new-citizenship-law-and-its-anti-secular-implications.

Knott, Eleanor. 2019. "Beyond the Field: Ethics after Fieldwork in Politically Dynamic Contexts." *Perspectives on Politics* 17 (1): 140–53.

Miljanovic, Morana, Vasilis Ververis, and Kersti Ruth Wissenbach. 2020. "Tools and Tactics for Data Protection Before, During, and After Fieldwork." In *Safer Field Research in the Social Sciences: A Guide to Human and Digital Security in Hostile Environments*, ed. Jannis Grimm, Kevin Koehler, Ellen M. Lust, Ilyas Saliba, and Isabell Schierenbeck, 108–25. London: SAGE Publications.

Morgenbesser, Lee, and Meredith L. Weiss. 2018. "Survive and Thrive: Field Research in Authoritarian Southeast Asia." *Asian Studies Review* 42 (3): 385–403.

Murgia, Madhumita. 2019. "Who's Using Your Face? The Ugly Truth About Facial Recognition." *Financial Times*, September 18. www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e.

Parkinson, Sarah Elizabeth, and Elisabeth Jean Wood. 2015. "Transparency in Intensive Research on Violence: Ethical Dilemmas and Unforeseen Consequences." *Qualitative & Multi-Method Research* 13 (1): 22–27.

Wackenhut, Arne F. 2018. "Ethical Considerations and Dilemmas Before, During, and After Fieldwork in Less-Democratic Contexts: Some Reflections from Post-Uprising Egypt." *The American Sociologist* 49 (2): 242–57.

https://doi.org/10.1017/S1049096522000543 Published online by Cambridge University Press