Taming Abundance: Doing Digital Archival Research (as Political Scientists)

Diana S. Kim, Georgetown University, USA

ABSTRACT Political scientists are increasingly using digitized documents from archives. This article is a practical introduction to doing digital archival research. First, it explains when and why political scientists use evidence from archival research. Second, it argues that the remote accessibility of digitized records provides new opportunities for comparative and transnational research. However, digital archival research also risks aggravating five types of biases that pose challenges for qualitative, quantitative, interpretive, and mixed-methods research: survival, transfer, digitization, and reinforcement bias at the level of record collection and source bias at the level of record creation. Third, this article offers concrete strategies for anticipating and mitigating these biases by walking readers through the experience of entering, being in, and leaving an archive, while also underscoring the importance of learning the structure of an archive. The article concludes by addressing the ethical implications to archival research as a type of field research for political scientists.

Among political scientists, the practice of using digitized documents from archives has become increasingly common. This article is a practical introduction to doing digital archival research. First, it explains when and why political scientists use evidence based on archival research. Second, it argues that the remote accessibility of digitized records provides new opportunities for comparative and transnational research. However, digital archival research also risks aggravating five types of biases that pose challenges for qualitative, quantitative, interpretive, and mixed-methods research: survival, transfer, digitization, and reinforcement bias at the level of record collection and source bias at the level of record creation. Third, this article offers concrete strategies for anticipating and mitigating these biases by walking readers through the experience of entering, being in, and leaving an archive, while also underscoring the importance of learning the structure of an archive. The article concludes by addressing the ethical implications to archival research as a type of field research for political scientists.

© The Author(s), 2022. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.
This article is a practical introduction to doing digital archival research.
Political scientists have always relied on humanistic solutions to problems of abundance in archival research.

How may political scientists doing archival research anticipate and address challenges of excess abundance from digitized records? The following subsections offer several concrete strategies that commonly center on ways to tacitly learn the structure of an archive.

Before Entering an Archive

Imagine that you are planning to travel to a new country. Learn the basic language that people at the archives use. Provenance (also known as respect des fonds) refers to the original creator of a record and its history ownership. It is a type of principle for arranging records in a way that preserves their integrity based on how they originated. It also informs the practice of original order, by which archives maintain records according to how creators arranged them originally. This is why records are not necessarily found in their original order. It also informs the practice of provenance, by which archives maintain records according to how creators arranged them originally. This is why records are not necessarily found chronologically, alphabetically, or according to geography. The organizing categories are those created by the individual, family, or agency from whom the archive acquired the documents.

Many archives arrange records according to a hierarchy. Collections are a general grouping of records that do not necessarily share the same provenance. Within a collection, the highest level of description is a fonds (or “record group”) in which records share provenance. A fonds is subdivided into series (and subseries), which are further subdivided into files. The lowest level of the hierarchy is an item, which is a record that is indivisible. The item-level record is what we usually understand as an archival document—the piece(s) of paper for a surveillance file of an individual, a court-proceedings transcript, or a tax record (figure 1).

Finding aids are one of the most important tools for navigating an archive (figure 2). They are detailed inventories of the records in a collection, containing metadata of a collection’s provenance, summary of contents and organization, administrative history and biographical notes, and size (e.g., number of boxes and linear feet of records). Some finding aids may include a file-by-file, item-by-item list of the collection’s contents. An index (or catalogue) is a list of records with shelf/call numbers. Research guides provide descriptive explanations of how to explore an archive’s holdings and often are written by an archivist or subject specialist. Think of each tool as a different genre of storytelling about an archive. Finding aids and indexes are often cryptic and not necessarily meant to be read from cover to cover. Rather, consult them selectively. Often hidden within their flat prose is invaluable contextual information about a collection. Research guides are more reader friendly, with rich narratives that should be consulted discerningly, not least because they are the products of another’s interpretation of an infinite archive.

Born-digital records are items that are created originally in digital form, such as emails, social media posts, and other types of electronic records. Digitized records generally refer to scanned copies of an original analog record; they are a type of access derivative. Just as paper-based records are fragile and can experience wear over time, digitized and born-digital records also face risks of data degradation (i.e., “bit rot”), losses in the process of transcoding and compression, and obsolescent formats.

Entering an Archive

Now you are at the archive, in the (virtual) reading room (figures 3 and 4). How do you begin to find documents? For political scientists in search of theory-driven evidence, a first helpful step is to develop a list of search keywords relating to their research question and tentative argument: X causes Y; A influences B. What are your concepts, words, and proper nouns relating to X and Y, A and B?

A surprising amount of archival research in the twenty-first century, whether in person or remote access, is time spent doing reiterative keyword searches. The search box, whether on an archive’s closed intranet terminal or public website, always mediates access.

However, digital archival research is not a Google search. To effectively use the search box, consider creating a two-layered set of keywords: (1) words that reflect how the archive labels and categorizes records relating to your X/A and Y/B; and (2) words in

---

**Figure 1**

Diagram of Levels of Archival Arrangement

---

**Figure 2**

Diagram of Levels of Archival Arrangement

context—that is, what past actors and institutions would have called your X/A and Y/B. To figure out the former, consult a finding aid, research guide, or index. To ascertain the latter, consult a seminal history or empirical study relating to your research.

For instance, suppose your X/A relates to street-level bureaucrats. The concept itself is an academic term of art. Which alternative words might capture the presence of those actors in the archive? Common sense may suggest “local administrators” and “municipal officials.”

Now think back to Lipsky’s (1980, 17–18) canonical study of street-level bureaucrats in the United States, which refers to them as “public service workers,” “public employees,” and “low-level workers.” Add these three keywords to your list. Perhaps urban Pakistan in the 2000s is your context; Hull’s (2013, 57–59) Government of Paper guides you toward this word: “clerks.” Perhaps the historical context of the early-twentieth-century British empire is relevant: “district officers” abound in Lugard’s (1922) The Dual Mandate in British Tropical Africa and Mamdani’s (1996) Citizen and Subject. Now turn to your Y/B. It includes the word “opium.” A digitized finding aid for the India Office Records at the British Library indicates that “opium” is cross-referenced with terms such as “Abkaree” and “Separate Revenue.” You already are starting to identify the controlled vocabulary of the archive, which has grouped together records according to their provenance of the British Indian colonial government for Burma’s Excise Department (Kim 2020).

Archival research is a reiterative process of discovery. By refining a list of search keywords rooted in both historical and archival contexts, political scientists learn tacitly about how “their” documents fit within the broader structure of an archive’s records, and they are better able to identify survival and transfer bias. Reiteration is an investigative process, not dull repetition, especially in a digital archive. Allow yourself to be distracted, especially by unexpected words, unfamiliar concepts, and odd proper nouns that pop up while scrolling and browsing. These are the moments when chance encounters that lead to new discoveries may occur.

### Being in an Archive

There is something both exhilarating and disorienting about being in an archive, physically or virtually. You have just found a 140-page archival document that seems promising, and there are so many more. What do you do? Taking notes systematically and organizing them according to original order are two seemingly mundane yet powerful strategies for addressing challenges of excess abundance.

First, design a consistent template for taking notes on each item that you consult.11 There is no right or wrong approach to notetaking. The template only needs to be one that can be...
Figure 3
In-Person Reading Room at France’s National Archives for Overseas Territories (Aix-en-Provence)

Source: Photograph by the author.
repeated over and over again. A minimalist may include only the file’s call number, title, and a brief summary of its contents. A maximalist may add the date accessed, date of original creation, copyright restrictions, a more detailed item-by-item description of its contents, and transcribed notes. Consistent and systematic notetaking at the item level is tedious and, at first, time-consuming. However, it becomes a habit that saves time in the long run. Crucially, it establishes a cumulative record of not only what types of records and information you selected to include in the final analysis but also what was not incorporated, which helps to mitigate reinforcement bias.

Second, systematic notetaking goes hand in hand with systematically organizing those notes. Consider mimicking the original order of the archive when storing notes as well as digitized copies of original documents. For instance, your 140-page document is from the UK National Archives in Kew. You located a digitized copy online and the original reference number is CO/885/1/20. First, create four nested files. Then, under “UK National Archives, Kew (digital),” create a second file labeled CO/885, followed by a third file labeled CO/885/1, and finally “20.” You have just replicated the record hierarchy for this specific document: Item No. 20 in the Miscellaneous Files (1) in the War and Colonial Department and Colonial Office Series (885), within the Colonial Office Records (CO) (figure 5).

Leaving an Archive
A realization hits you as you leave an archive, whether walking out of the building or closing your web browser for the day. The documents that you consulted and the records that you accessed bear traces of the lives of others. Inevitably, the archival research you have just done is an encounter with people in history.

Digital archival research is inextricably tied to ethical considerations for political scientists. Survival, transfer, digitization, and reinforcement biases are products of how archives are not impartial repositories but rather institutions shaped by power, politics, and privilege. Remote accessibility makes it easier to forget how selective and partial an archive’s holdings are, not least because it eliminates the many inconveniences that remind researchers of its arbitrariness and incompleteness. Thus, there is a greater risk of overrepresenting the coherence of past events, processes, and human experiences because it is easier to presume the digitized available documents capture a greater share of historical reality than actually warranted. The technology of digitization also generates new considerations that bring the ethics of historical representation squarely into the ambit of political science’s theory-driven evidence seeking. OCR errors can result in the erasure of an individual’s trace in the archives and, conversely, digitizing microlevel data may inadvertently “find” people with vexed histories unknown to their descendants. The boundaries of
copyrights also are blurry for digitized records extending to photographs. A keen sensitivity to these issues animates emerging scholarship on archival and historical research for political science (American Political Science Association 2019, 2021; Balcells and Sullivan 2018). To bring ethical considerations to the forefront, a basic and necessary question for researchers to ponder is: Why do we choose to do what we do with archival records? Although there is no right answer, there are many different types of thoughtful responses. For some scholars, there is an impulse of social justice, of advocacy on behalf of the weaker, the voiceless—not least to rescue them from “the enormous condescension of posterity” (Thompson 1963, 12). Other scholars may seek not to speak for other people. “The intention here isn’t anything as miraculous as recovering the lives of the enslaved or redeeming the dead,” Hartman (2008, 11) wrote of her approach to the archives of the eighteenth-century transatlantic slave trade. Rather, she explained, it is about “laboring to paint as full a picture of the lives of the captives as possible” (Hartman 2008, 11). Different epistemologies may counsel archival research as a way to “better understand how local history and context can be leveraged to inform the design of better policy” or, alternatively, to gain a richer menu of counterfactual empirical realities for rigorous social-scientific inquiry (Fouka 2020; Nunn 2020, 1). In doing archival research, choices are already being made, with stakes that are amplified in the process of remotely accessing digitized records. Digital archival research gains ethical import when political scientists are able to recognize and explicitly articulate these choices.

ACKNOWLEDGMENTS

I am indebted to Jennifer Cyr, Daragh Grant, Diana Kapiszewski, N. M. Kim, Ian Kume kawa, Jean Lachapelle, Lauren MacLean, Kate McNamara, Emma Rothschild, Kyle Shen, Yuuki Tajima, and Htet Thiha Zaw for illuminating conversations and constructive comments. I also am grateful to participants in the modules for Designing and Conducting Field Research and Interpretation and History at the 2021 Summer Institute for Qualitative and Multi-Method Research (IQMR).

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S104909652100192X.

CONFLICTS OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

NOTES

1. On epistemological and methodological differences between social scientists and historians when approaching archival evidence, see Gaddis (2002), Lemercier and Zalc (2019), and Sewell (2005). On archival methods in field research for political scientists, see Kapiszewski, MacLean, and Read (2023). On how COVID-19-related interruptions are reshaping field research, see MacLean et al. (2020).
REFERENCES


Lugard, Frederik. 1922. The Dual Mandate in British Tropical Africa. London: Blackwood and Sons.


The Profession: Doing Digital Archival Research (as Political Scientists)

2. Here, I refer mainly to official archives in the form of material records, both textual and non textual, that either are preserved at the producing institution or collected by an external repository. This is only one among a vast variety of entities that scholars recognize as archives, ranging from informally kept records to nontangible embodiments of collective memory, past ideas, and lived experiences.

3. On survival and transfer bias, see Lee (2017, 5-6) and, more generally, on how biased sampling from archival research can affect hypothesis testing.

4. On selection bias as a type of source bias, see Lustick (1996). On problems of scale that influence biases, see Kumekawa (Forthcoming).

5. For exemplars of such humanistic approaches among historians, see Farge (2015) on the role of tacit learning and tactile experience in archival research; Trivellato (2000) on recovering a "lost canon" through deep contextualized readings of sources identified through data-mining tools; and Rothschild (2013) on reconstructing complex transnational family and social ties using large-scale network visualization, combined with micro-histories of individuals. For an illuminating example of these visualizations, see www.infinitehistory.org/en/networks.html.

6. For valuable "how-to" guides detailing step-by-step processes from identifying archival sites to gaining access and technologies for digitizing records, to writing and analysis, see Redman (2013) and Abbott (2014).

7. For more on provenance and original order, see Schellenberg (1951). On the modern history of provenance as an archival principle, see Sweeney (2008).

8. For valuable guides detailing step-by-step processes from identifying archival sites to gaining access and technologies for digitizing records, to writing and analysis, see Redman (2013) and Abbott (2014).

9. Of the first virtual reading rooms for a US academic repository was established for Richard Rorty’s papers at the University of California, Irvine (Liptsey 2009).


11. See online appendix 1 for a sample template.


