

The Library of Babel Problem: Hypothesis Testing With Archival Sources

Alexander Lee*

November 24, 2017

Abstract

The use of archival sources is understood to be an important tool research tool, but the problems involved have rarely been distinguished from the broader difficulties of causal inference with qualitative sources. Attempts to use archival material for hypothesis testing, as opposed to description or theory development, are troubled by the large size and often opaque structure of archives; factors which lead to misinterpretations of evidence and a tendency to confirm the author's expectations. This paper discusses common features of archival materials, shows how they can compound traditional research design problems. It then proposes a set of best practices for avoiding these problems, most notably the use of strong and explicit sampling procedures.

Key words: Archives, Sample Selection

Word Count: 3648

*Department of Political Science, University of Rochester

1 Introduction

Archival sources, like other types of qualitative and textual evidence, present substantial challenges in their interpretation and the making of causal inferences. Sources often are unreliable, disagree with one another, reflect the biases of their creators. The voluminous literature on qualitative methods has discussed how to deal with these problems. Advice on archival methods targeted to political scientists tends to recommend that they follow historical best practices (Trachtenberg, 2009). Similarly, more general work on process tracing and case study research emphasizes finding solutions to interpretive problems and the difficulties in making reliable causal statements based on evidence (Bennett and Checkel, 2014; Kreuzer, 2010; George and Bennett, 2005; Beach and Pedersen, 2013).

However, archival material presents several additional unique and poorly understood challenges. Archives often feature a very large amount of poorly organized material, making it difficult for scholars to access the full extent of the material produced by the institution. In addition, the preserved information is often unrepresentative of that originally produced. and the patterns of missingness are not always obvious. These shortcomings present temptations for scholars to make inferences based on a biased subsample of the archive.

For many of the purposes for which archival material is conventionally used, these problems are relatively benign. Scholars who aim to explain specific outcomes, or describe events, may avoid these problems if they confine themselves to small and well-preserved portions of the archival record. However, in many cases, political scientists seek to use archival material to develop or test causal hypotheses. Historical techniques, which emphasize an inductive and descriptive approach to archival material, can produce biases when used to test these sorts of hypotheses, most commonly bias towards confirmation. At their worst, the combination of inductive historical techniques with a strongly held existing theory can lead scholars to use archival sources as a mine for confirmatory quotes without considering the incompleteness and complexity of the source material.

To address these problems, it is necessary to analyze archival material in a way

that integrates its unique features into standard methods of hypothesis testing. This paper highlights three methods that can help in the analysis and presentation of archival evidence, and are widely used in non-archival studies. Firstly, and most importantly, scholars should develop an explicit *a priori* rule for what types of material will and will not be analyzed. This reduces the danger (encouraged by both traditional archival filing systems and modern content management software) that they only analyze material that supports their existing ideas, and enables them to be more aware of underlying patterns of missingness within the archive. Secondly, scholars, following standard historical practice, should become familiar with the internal procedures and personalities of the institution that produced a particular set of sources, so as to be familiar with both the biases inherent in the written material, and what parts of the archive are appropriate for study. Finally, scholars should acquaint their readers with the exact procedures that they used to gather material, enabling readers to critique these procedures and allowing the readers to understand how specific contentions are backed by evidence.

2 Common Problems in The Use of Archives

2.1 Definitions

Archives are accumulations of unpublished historical records, usually of an institution or individual. The term also refers to the specialized facilities in which these accumulations are held, though some agencies may continue to maintain their archives directly, and many archival facilities contain large amounts of published material.

Some social scientists casually refer to all source material on historical topics as archival, even published secondary work. This usage is misleading, since as part of the process of publication much factual material is inevitably excluded. While published primary source material is often much easier to access, and more clearly presented, than archival material, it is usually less rich, and more biased, than unpublished papers. Published secondary source material adds to this selection effect the interpretive biases of its authors, making any attempt to interpret them as historical source material more difficult (Lustick, 1996).

2.2 Confirmation Bias and the Library of Babel

One of the most common features of archives is their vast size relative to the interpretive capacities of scholars. The American National Archives contains over 10 billion pages of material, while even the humble Guatemalan Police archives total some 10 million pages. Even if the topic is narrowed, the capacity of institutions to produce paper runs well ahead of the capacity of scholars to understand it. A scholar interested in the foreign policy of India between 1947 and 1955 has 42,654 files to choose from, the majority handwritten. The problem of size is compounded by problems of institutional organization. In most archives, files can only be requested in finite amounts through a time-consuming procedure, browsing of material is forbidden, and indexes and other finding aids are often fragmentary and cryptic. Scholars must thus invest considerable time in obtaining relevant material before they are able to analyze it.

Some selection of what material to view is thus imperative in any type of archival work. This selection is, however, very dangerous from a research design perspective, since any selection of cases may potentially lead to biased findings. While bias resulting from case selection is a well-known danger in all forms of social scientific inquiry, it is particularly insidious in the archival context, due the opacity of archives to both scholars and those accessing their work. A scholar selecting countries for detailed case studies or congressmen for interviews is aware in a general way of the characteristics of the universe of countries and congressmen, and can thus select a representative and theoretically relevant sample, or at least be aware of the potential biases of the sample they do obtain. The selection criteria are ideally explained in the work itself, where they can be assessed by readers. An archival scholar, by contrast, may have little sense of what the “universe” of files may look like, and his or her reader even less. A scholar may thus fasten upon a particular subsection of material that supports her hypothesis, without having any sense of its wider validity. Given the vast size of archives, most contain material that can be used to support a wide variety of plausible hypotheses.

While scholars doing tightly focused or descriptive work, also face the problem of choosing what archival material to read, they are less troubled with bias from file selection. To the extent that scholars choose topics on which the archival material

is definite and finite, they may be able to read all the available material on the topic, reducing the probability that they will base their conclusions on an unrepresentative portion of the source base.

However, many political scientists begin their archival research with a strong, clearly articulated hypotheses about the findings they will encounter. While such hypotheses are a key part of social scientific inquiry, their existence means that scholars able to process only a small amount of information may consciously or unconsciously choose information that supports their existing ideas.

To see what such unconscious selection bias might look like, consider again the case of a hypothetical scholar who wishes to understand the effect of colonialism on religious identity. Let us imagine that our scholar has a strong belief that colonial-era policies are a key cause of religious tensions in this area. After searching an index for entries on “religion,” he or she encounters a wide variety of materials showing colonial involvement in religious affairs—missionary education, subsidized Muslim ulema, job quotas etc. From this he or she may conclude that the colonial state was heavily involved in the promotion of religious difference, and return home. While such a perspective may be correct, it may ignore archival material in which the colonial government may promote alternate types of cultural differences, or material in which religions considerations are simply not present. Such archival selection is normally difficult to detect, since readers typically have little idea of the universe of files from which scholars are selecting their evidence.

The easier availability of electronic indexes, digitized archives and content management software all promise to remove much of the drudgery from archival work, but may at the same time compound the problem of selective reading, by making it easier for scholars to isolate files, or even individuals sentences, that support their thesis, and making it less likely that they will encounter alternate perspectives. [Neitzel and Welzer \(2013\)](#), for instance, use text analysis of transcribed POW conversations to assess the involvement of German soldiers in atrocities during the Second World War. While their technique finds a shocking number of incidents in which atrocities were discussed, it is difficult to access their status as evidence, since the authors have not analyzed the much larger number of “ordinary” conversations.

2.3 Missing Data

Not only are scholars unable to read or code every file in an archive, but the files in an archive are inevitably a subset of the files produced by the institution, and represent an even smaller subset of the information available to the institution. In archival contexts, as in other types of social scientific inquiry, missing data can produce bias in both qualitative and quantitative inference. This bias takes three primary forms:

1. “Survival Bias” is a product of the long periods of time, and the occasionally uncertain environments, that intervene between the producers of material and the scholar’s desk. During this time, there are abundant opportunities for material to be lost, stolen or destroyed. Sometimes these factors are environmental, like the termite infestation at the West Bengal State Archives that has severely eroded our ability to understand early colonial administration in India. Sometimes they are human, as in the USAAF’s 1945 air raid on Potsdam, which destroyed the archives of the German Army and severely limited study of the German role in the First World War. On a lower level, archival institutions, especially in the developing world, suffer low level leakage of material: Files are checked out and not returned (and occasionally found a few decades later in a professor’s attic), files are returned but misplaced, files stored in the front of boxes become flaky and illegible, and maps are taken for sale to dealers.

While these sources of bias may seem random and ignorable, this is not always the case. At every stage of the preservation process, chances favor the survival of material considered important. The Potsdam Air raid, for instance, did not affect certain classes of papers that had been moved away for safekeeping, while the WBSA staff appears to have kept the termites away from files with obvious relevance to left-wing political activism. Similarly, institutions and regions which remain stable over time tend to produce more complete archival records, since wars, revolutions and reorganizations are obvious opportunities for archives to be destroyed or neglected.

2. “Transfer Bias” is a product of the institutional gap between the individuals or institutions that produce material and the archival institutions that store and disseminate them. Producing institutions may be reluctant to transfer material that they deem embarrassing or sensitive, which they retain in their “active” files.” Britain’s Security Service, for instance, has released no material since the 1950s, while many

US presidential libraries embargo material that they consider sensitive. Even assuming perfect goodwill, the period of transfer often becomes a focal point for the discarding and dispersion of material.

3. “Source Bias” is the most common and problematic of the three types of bias. It stems from the fact that government agencies collect and record certain types of knowledge in preference to others. For obvious reasons, governments tend record information about their own activities rather than presenting a record of society as whole: We know far more, for instance, about taxation in the Middle Ages than we do about gender relations. Controversial matters, for instance, tend to produce multiple submissions, appeals to higher authority, and thus a much larger paper record than “routine matters.” Divorces, classically, are better recorded than happy marriages.

What makes these sources of bias particularly damaging for hypothesis testing is that the exact extent and direction of the bias is difficult to determine easily. While a scholar analyzing a survey dataset, for instance, may know exactly which respondents failed to answer a particular question, and an interviewer of elites may know which individuals avoided his requests, archives sometimes fail to make it obvious what types of information have been lost. Faced with such a situation, social scientists will often cite whatever information is available as the “archival” truth without considering its place within the larger historical record.

2.4 Different Goals in the Use of Archives

As with other types of qualitative and quantitative sources, scholars may have many different goals in the interpretation of archival material. These may include description of particular events, the generation of broadly valid causal hypotheses, and the testing of those hypotheses (Beach and Pedersen, 2013). Scholars pursuing the last of these goals, however, are most vulnerable to biases in the archives. To the extent that they seek to make broadly relevant causal claims, they must take cognizance of the whole universe of potential data, even when it is not available.

Obviously, even if a scholar had every possible document available and was granted infinite time to read all of them, his or her task would not be done. She or he would have to decide the correct interpretation of individual documents that are often ambiguous,

one sided, or cryptic. She or he would then have to decide how to “scale” these interpretations into a larger argument about the material and the truth of the hypotheses. Neither of these tasks is trivial, and a voluminous literature has grown up to advise scholars on how to accomplish them (Bennett and Checkel, 2014; Kreuzer, 2010; George and Bennett, 2005; Beach and Pedersen, 2013). This is particularly true since neither problem is unique to archives: Many kinds of social scientific information require interpretation before they can become useful for making causal arguments.

3 Best Practices for Archival Social Science

3.1 Understanding How the Data Was Created

A scholar seeking to use archival material to test a hypothesis should first emulate existing historical best practice by making herself familiar with how the material was presented and preserved. Such familiarization is essential in order to collect the most relevant information from the archive, understanding the biases inherent in the availability or non-availability of material, and the internal conflicts within the organization.

At its most basic, this background research involves identifying, through the finding aids in the archives and secondary sources, what institution or individual created the records and how they were organized. In some cases this is simple: The US Supreme Court, for instance, has maintained a fairly constant organizational profile for over two centuries. The responsibility for enforcing federal alcohol tax laws, by contrast, has been shuffled through a bewildering number of agencies in the same period, and these agencies have been reorganized and expanded in accordance with political convenience and administrative fashion. Generalizations such as “Federal Authorities” may conceal considerable variation in institutional cultural and incentive structure.

After understanding the structure of the institution, the scholar should try to understand the internal flow of paper. This means stepping back from viewing archives as masses of sources, and seeing them as their creators saw them—as individual cases flowing from desk to desk, or a pile of letters to be responded to. Some of the important questions in such an analysis are who initiates matters, who receives information, and

who makes key decisions. In some organizations, peripheral agents may be focused on implementing central policy dictates, while in others the center contents itself with providing resources for field agents. Understanding this flow is important in choosing which documents to focus on. While it may be tempting to quote an impassioned memorandum on one side of a question (particularly if it agrees with the hypothesis) this memorandum may be misleading if it was produced by someone with little effect on decision-making.

3.2 Creating a Sampling Frame

An important step towards unbiased archival hypothesis testing, is to make explicit at the beginning of the main analysis what material will be reviewed and what will not. All data within this frame should be reviewed, and become part of the conclusion, either qualitatively or quantitatively. Such planning serves a number of important purposes. While it does not increase the amount of material that can be read, it guarantees that the scholar make explicit, at least to himself, the criteria by which the material to be read is selected. This reduces the potential for the author to shape the findings by choosing what files to read. Similarly, while an explicit sampling frame cannot recreate the data destroyed by the various types of archival bias, it can make more obvious to the scholar what is missing and why.

In development of sampling frame, scholars should be guided by the structure of the archival material, with which they will be familiar after the exploratory study recommended above. A scholar might choose to read all files from a specific department for a specific period, or all of a specific class of document, or all cases under consideration at a specific time. To take some examples, *citetblaydes2013compliance* collects all instances of a specific type of document, the school report on political reliability, for a given year, [Lee \(2011\)](#) uses as observations all Indian Nationalists listed as political suspects in a given year, and [Hainmueller and Hangartner \(2013\)](#) use all immigration cases decided by a Swiss Canton in a given time period. While in all these cases some elements of the sampling frame were determined by data availability (the specific year, the specific Canton) the inclusion of individual files in the data frame was determined by the structure of the archive itself.

After the main discussion, a scholar may still have many questions about the data that might not be answered within a narrow sampling frame. This might include questions about the applicability of the finding to other contexts, or the situation in areas or time periods that are missing in the main data. These questions may become the basis for the construction of additional sampling frames and further analysis. Even if the additional material is not useful for a sustained analysis, a scholar may wish to consider a brief examination of the additional material to see if it is consistent with the findings from the main analysis.

3.3 Providing Context for the Reader

The understanding of bias and the construction of sampling frames are of little use to scholars if they are not communicated to readers. While scholars are habituated to sharing with readers the details of the construction of quantitative datasets, or the organization of anthropological research, they are unwilling to share similar levels of information about their archival research, making their claims sometimes less than fully convincing. Compare two different ways of presenting the same set of hypothetical facts.

(1) To examine the motivations for this unprecedented expansion of primary education, I examined the files of the educational department. The archival evidence shows that government officials saw the expansion primary education as a method for reinforcing national unity, rather than as a way of enhancing economic development.

(2) To examine the motivations for this unprecedented expansion of primary education, I examined all files produced by the construction subsection of the education department between 1949 and 1965, a total of 6,000 pages of material. Overwhelmingly, they mention as the motivation for construction the need to tie remote villages into a national community, rather than the economic gains from education.

While both these statements present the same facts, the second is far more convincing, since it gives the reader a basis for judging the solid evidentiary basis on which the statement rests.

The gains to openness can be made even larger when scholars are willing to acknowledge that their sources are not always unanimous and discuss the extent of the disagreement:

(3) To examine the motivations for this unprecedented expansion of primary education, I examined all files produced by the construction subsection of the education department

between 1949 and 1965, a total of 6,000 pages of material. Overwhelmingly, they mention as the motivation for construction the need to tie remote villages into a national community, rather than the economic gains from education. While national identity value of school construction is mentioned in dozens of instances, on only two occasions are the economic benefits of education mentioned.

4 Conclusion

Archival material can potentially be a valuable source for political scientists, allowing them to test hypotheses about institutional behavior and information gathering that are inscrutable in a contemporary setting. However, the size, disorganization, and non-comprehensiveness of most archives create problems for causal inference, especially selection and missing data. This paper proposes a set of simple suggestions for dealing with these problems: An understanding of the biases inherent in the sources, the development of an explicit frame for reading and analyzing the material, and sharing these methods and findings with readers.

While certain of the approaches discussed here, such as an understanding of the producing institution, are useful in a wide variety of contexts, these techniques are not intended to be a universal prescription for archival research. Scholars interested in process tracing and description may find the construction of a hard sampling frame to be unduly burdensome and narrowing. However, these techniques are very appropriate for those who seek to use archives to test existing theories. By enabling causal claims to be made more credibly, broadening the influence of archival evidence within the discipline, and improving our understanding of the inner workings of political institutions.

Bibliography

Beach, Derek and Rasmus Brun Pedersen. 2013. *Process-tracing methods: Foundations and guidelines*. University of Michigan Press.

Bennett, Andrew and Jeffrey T Checkel. 2014. *Process tracing*. Cambridge University Press.

George, Alexander L and Andrew Bennett. 2005. *Case studies and theory development in the social sciences*. Mit Press.

- Hainmueller, Jens and Dominik Hangartner. 2013. "Who gets a swiss passport? a natural experiment in immigrant discrimination." *American political science review* 107(01):159–187.
- Kreuzer, Marcus. 2010. "Historical knowledge and quantitative analysis: The case of the origins of proportional representation." *American Political Science Review* 104(02):369–392.
- Lee, Alexander. 2011. "Who Becomes a Terrorist?: Poverty, Education, and the Origins of Political Violence." *World Politics* 63(02):203–245.
- Lustick, Ian S. 1996. "History, historiography, and political science: Multiple historical records and the problem of selection bias." *American Political Science Review* pp. 605–618.
- Neitzel, Sönke and Harald Welzer. 2013. *Soldaten: On Fighting, Killing, and Dying: The Secret World War II Transcripts of German POWs*. Knopf.
- Trachtenberg, Marc. 2009. *The craft of international history: A guide to method*. Princeton University Press.