Online Appendix for "Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them"

Jonathan Grossman and Ami Pedahzur

1. CONTENT AND CITATION ANALYSIS OF POLITICAL SCIENCE ARTICLES ABOUT BIG DATA

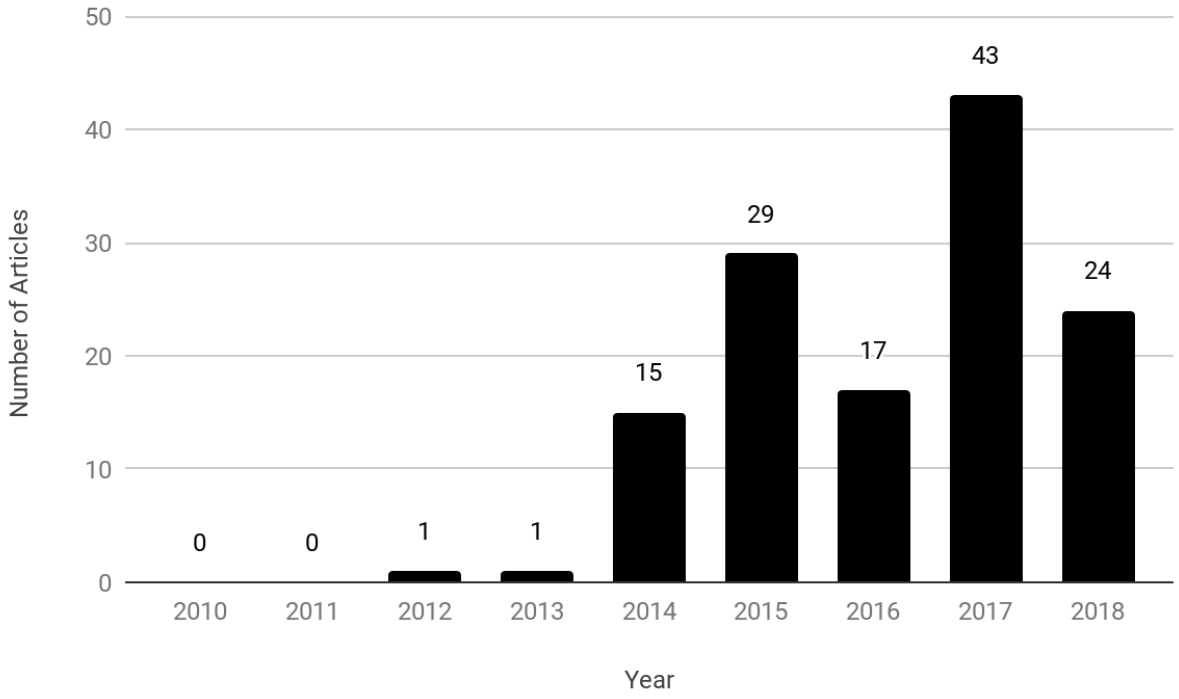
To analyze the state of the discipline, we examined bibliometric data from Clarivate Analytics' *Web of Science* (WoS) academic index. Despite the known shortcomings of WoS, which is a proprietary database whose inclusion criteria are not entirely transparent,[1] we consider it among the best available sources for citation data, especially when exploring the mainstream bibliography of a discipline.[2]

We searched the *Web of Science* on 1 January 2019 for papers whose topics included the expression "big data." We downloaded all *research articles* (that is, not conference proceedings, reviews, editorials, etc.) that were published in journals whose topic was categorized by WoS as political science (n=87), international relations (n=35), and public administration (n=36).[3] All in all, there are 133 such articles, 132 of which are in English and one is in Spanish, with the earliest article being published in 2012.[4] As Figure 1 shows, since 2014 there has been a marked increase, although not a dramatic or a steady one, in the number of political science articles about big data. Some of this growth was instigated by special issues: 30 articles were published in the

[1] Diana Hicks et al., "Bibliometrics: The Leiden Manifesto for Research Metrics," *Nature* 520, no. 7548 (April 2015): 429–431; Philippe Mongeon and Adèle Paul-Hus, "The Journal Coverage of Web of Science and Scopus: A Comparative Analysis," *Scientometrics* 106, no. 1 (January 2016): 213–228, at 218–19.
[2] Peter Marcus Kristensen, "International Relations at the End: A Sociological Autopsy," *International Studies Quarterly* 62, no. 2 (June 2018): 245–259, at 246–47.
[3] These numbers are not exclusive. The *Web of Science* may assign up to three categories to the same journal. Further, WoS periodically updates its indexing, constantly adding and removing articles and categorizations. Therefore, the numbers presented here are likely to change retroactively every once in a while.
[4] The full list of articles and all the variables in this analysis can be found at https://github.com/for-anonymous-review/Big-Data-in-Political-Science. The actual number of articles may be somewhat higher: a few political science and policy articles about big data were not included in the *Web of Science* search results for unknown reasons. The two examples that we were able to find using other search methods were Sandra González-Bailón, "Social Science in the Era of Big Data," *Policy & Internet* 5, no. 2 (June 2013): 147–160; Maureen A. Pirog, "Data Will Drive Innovation in Public Policy and Management Research in the Next Decade," *Journal of Policy Analysis and Management* 33, no. 2 (Spring 2014): 537–543.

same issue with at least two other articles from the database; 41 articles share the same volume with at least two other articles.

FIGURE 1: POLITICAL SCIENCE ARTICLES ABOUT BIG DATA, BY YEAR



Source: chart created with data from a Web of Science search conducted on 1 January 2019 with the following specifications: DOCUMENT TYPE = "Articles"; WEB OF SCIENCE CATEGORIES = "Political Science," "International Relations," and "Public Administration."

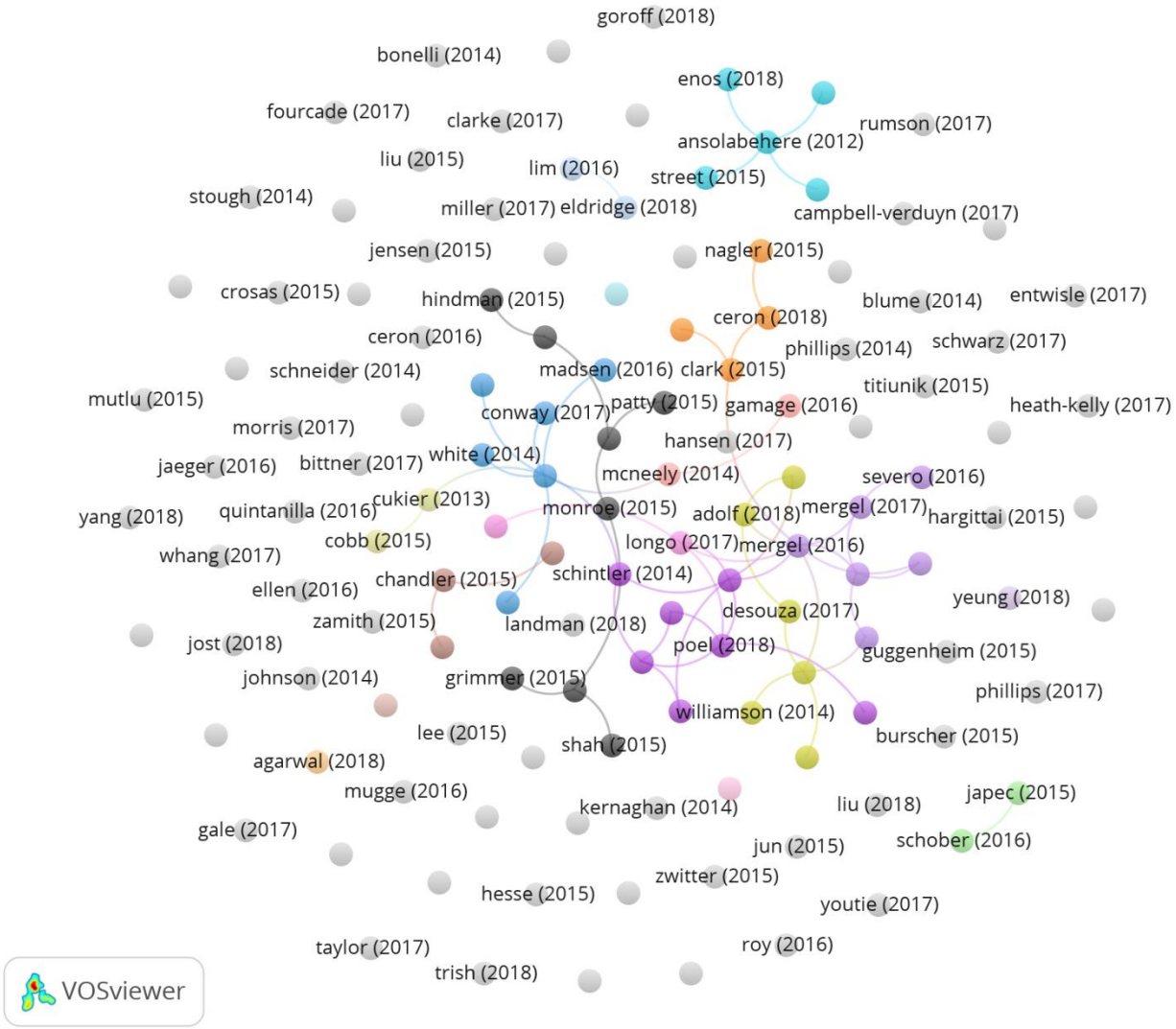
Next, we reviewed the content of each one of the 133 articles, both manually and with computer-assisted qualitative data analysis software (*NVivo*), and checked how many times and

in what parts of the article the term “big data” occurred. The analysis revealed that, in some cases, the term served as a buzzword (or, perhaps, as clickbait) to catch the attention of readers or journal editors without much actual discussion of big data. In sixteen cases, the expression ”big data” only occurred in the article title, abstract, list of references, or list of keywords (including keywords designated by the *Web of Science* when a journal failed to provide such a list). Fifteen articles mentioned big data only once; in 35 articles, the term occurred three times or less. Only 52 articles – less than 40 percent of the 133 articles – offered a definition of big data that was workable to some extent or that cited existing definitions.

A citation network analysis of the 133 articles (using VOSviewer software[5]) suggests that the discussion of big data in political science is highly fragmented. Less than half of the articles in the database are linked to each other through citation (Figure 2). This includes even articles that gathered a relatively high number of citations by papers not in the database. The few articles that cite one another are often those that were published in the same special issue or by the same journal. As a co-citation analysis (Figure 3) demonstrates, many of the 133 articles draw on a small pool of seminal works from other disciplines about big data, while seldom referring to the rather limited literature on big data in political science.

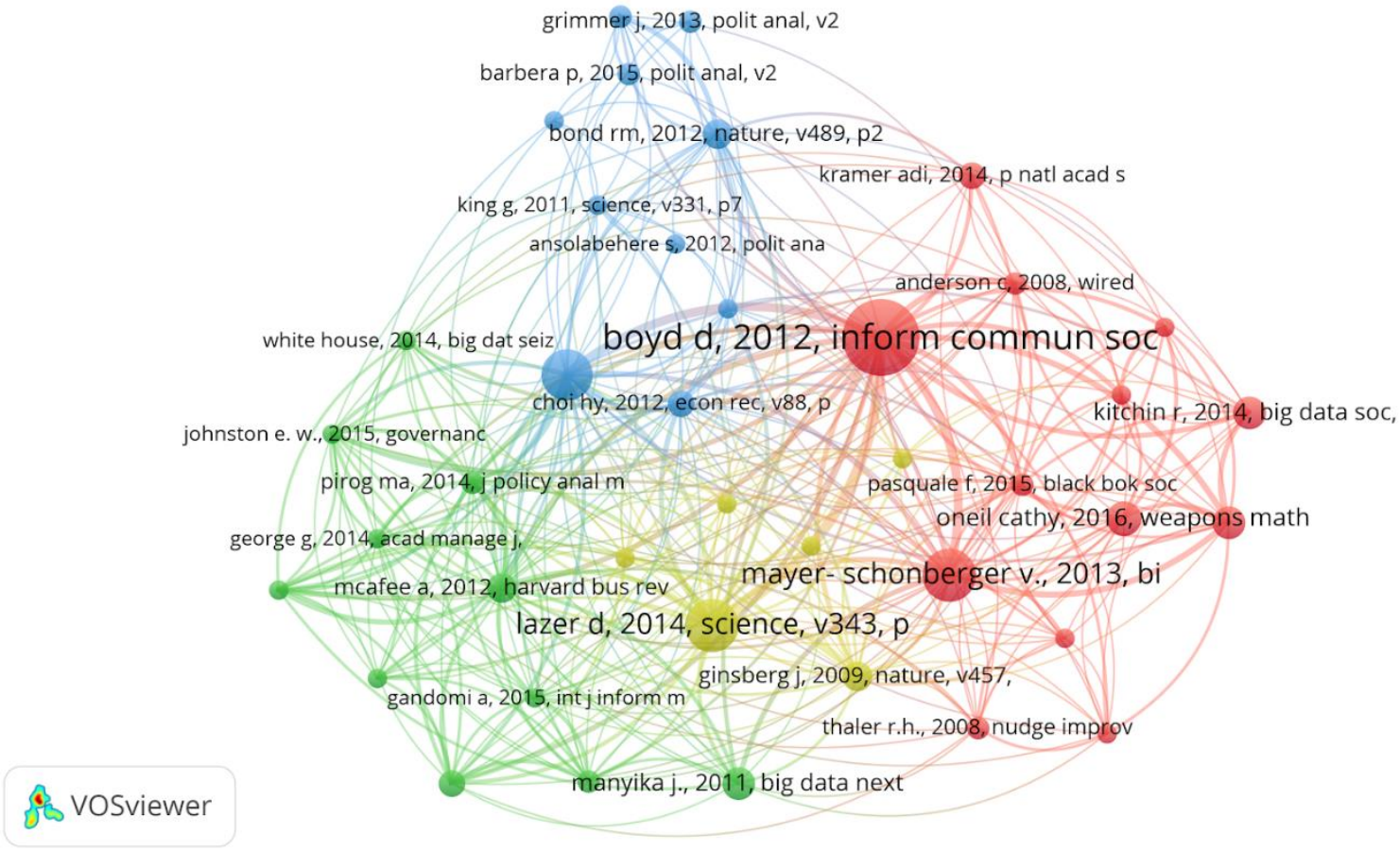
[5] http://www.vosviewer.com/. See Nees Jan van Eck and Ludo Waltman, “Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping,” *Scientometrics* 84, no. 2 (August 2010): 523–538.

FIGURE 2: A CITATION NETWORK OF POLITICAL SCIENCE ARTICLES ABOUT BIG DATA



Source: figure created using VOSviewer software and Web of Science bibliometric data (1 January 2019). A link between two nodes means that the more recent paper cites the earlier one. The size of each node represents the total number of citations that the paper has (according to WoS). The total number of citation links in the database is 59.

FIGURE 3: A CO-CITATION NETWORK OF POLITICAL SCIENCE ARTICLES ABOUT BIG DATA



Source: figure created using VOSviewer software and Web of Science bibliometric data (1 January 2019). Only papers that are cited by at least four articles in the database are shown (n=40). The cited works are represented by nodes. A link between two nodes means that at least one article in the database includes both articles in its list of references.

2. FINDING DATES, EVENTS, ACTORS, AND LOCATIONS IN UNSTRUCTURED BIG DATA SOURCES

When large collections of unstructured data are available, it would be impractical to carefully read each and every document, article, or book and to manually identify and record all the temporal units, actors, locations, and events that occur in the text. Unless one has a huge, devoted team of highly trained research assistants at one's command, such a task will take far too much time.

To address this issue, we propose an additional approach: *regular expressions*. This concept refers to common patterns in a text as well as the ability to recognize such patterns.[6] Using regular expressions requires some digital literacy skills but not necessarily any programming knowledge; to find patterns in a body of text, it is possible to use one of the many online resources that allow uploading or pasting text into a designated box and specifying the terms of the search.[7] However, scholars with a basic background in programming languages such as Python can automate the process of text recognition without much effort.[8] A more sophisticated option would be to use a learning algorithm that, through a process of trial and error, would eventually detect patterns of interest.

How can we use regular expressions to identify temporal units, events, actors, and locations in big unstructured data? Dates and times would be the easiest to find, as they have a unique textual pattern that is distinguishable from other parts of the text. The format of presentation may vary – for example, the same date can appear as "26 September 1959," Sep. 26,

[6] Jeffrey E. F. Friedl, *Mastering Regular Expressions*, 3rd ed. (Sebastapol, CA: O'Reilly, 2006), 4.
[7] For example, https://regexr.com/, accessed 19 August 2019.
[8] Al Sweigart, *Automate the Boring Stuff with Python: Practical Programming for Total Beginners* (San Francisco, CA: No Starch Press, 2015), chap. 7.

1959," "09/26/59," "26.9.1959," "1959-09-26," and so forth. In fact, the year may not appear at all. However, the number of possible formatting options is finite. By searching for all these patterns in an unstructured source, researchers can pinpoint the day or moment in which something that was worth reporting transpired. Based on their expertise and contextual knowledge, they can then decide which of those events are relevant to their questions and hence worthy of inclusion in the timeline. Once again, it is of the utmost importance that researchers document the exact search terms that they had entered as well as their inclusion criteria in order to be able to fix errors and mistakes or replicate the process at a later stage.

The names of places and people constitute a somewhat different category. While they appear to be an integral part of the text, they do have some characteristics (such as capitalization patterns and common suffixes) that can distinguish them from other parts of the text. Political scientists with a background in programming and knowledge in the field of natural language processing – or joint teams composed of political and computer scientists – may pursue additional automated approaches. For example, information extraction systems can mine text on the basis of preloaded dictionaries of common names or locations.[9] Further, they can carry out *named entity recognition* tasks, whereby the system automatically identifies entities such as date and time, people, locations, and organizations.[10] By writing such an algorithm or installing an available package, researchers can extract from the data a list of actors, locations, and activities that are candidates for inclusion in the codebook and hence in the analysis. While the named entity recognition approach involves a high degree of automation, it is still an auxiliary method. Researchers must decide which entities to include and how to carry out the analysis.

[9] Ronen Feldman and James Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge; New York: Cambridge University Press, 2007), 106–107.
[10] Ibid., 96–97; Gregor Wiedemann, *Text Mining for Qualitative Data Analysis in the Social Sciences* (New York: Springer Berlin Heidelberg, 2016), 33–34.

Alternatively, and more deductively, researchers can run a text search for specific words of interest. Many software packages, including free document readers, can perform advanced searches in multiple documents. Researchers who know what they are looking for can create a list of relevant search terms in all possible variations – for example, "Dwight Eisenhower," "Dwight D. Eisenhower," "Dwight David Eisenhower," "'Ike' Eisenhower," etc. Searching for these strings, the researchers can isolate the chunks of text in which they occur to read them more closely. Doing the same with verbs of interest (for example, "visit," "declare," "fight"), they may find information about events. Some software packages also offer more sophisticated search capabilities, such as stemming and automatic detection of synonyms. However, this method is only valid for unearthing additional information on known entities rather than discovering new ones. This limitation underlines, once again, the importance of background knowledge, context, regional expertise, theory, and clearly defined questions and hypotheses for causal inference in political science.

3. A STRUCTURED APPROACH TO UNSTRUCTURED BIG DATA

As we argue in the article, the key to dealing with unstructured big data is to impose order on chaos. This goal can be achieved by creating accurate timelines, that is, identifying relevant events, organizing them in a temporal order, and confirming the locations in which they took place, the identity of the actors that participated in them, and the interactions between these actors.

Thus, a research project that makes use of unstructured big data should be accompanied by a codebook that allows researchers to document all the observable elements relevant to their question in a structured and replicable manner. The most generic template for such a codebook is a spreadsheet that contains, in different columns, the name of an item of interest, information about it, and the instances in which the item occurs in the data.[11] This third column should allow researchers to instantly access the original, unstructured source in which the item is discussed. To this end, the creators of the database can copy and paste the relevant text excerpt into a matching cell, attach the data file itself (as a text, image, audio, or video document, if the computer program that the researchers use allows it), or provide a direct and permanent link to a copy of the source.

We propose utilizing the codebook to trace processes, establish causality, and rule out rival explanations. Since each part of a causal mechanism is "composed of entities that undertake activities,"[12] one needs, when analyzing these mechanisms, to identify participants unequivocally, record their precise whereabouts in specific periods, and trace their interactions in a way reminiscent of the board game *Clue* ("Colonel Mustard, in the ballroom, with a rope"). Thus, we recommend arranging the codebook by the following four categories: *temporal units*

[11] John W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 3rd ed. (Thousand Oaks, CA: Sage Publications, 2009), 188–189.
[12] Beach and Pedersen, *Process-Tracing Methods*, 49.

(when and for how long did the activity take place?), *events* (what was the nature of the activity?), *actors* (which individuals, groups, organizations, or other entities participated in the activity?), and *locations* (where did the activity happen?). Initially, the codebook is based on a preliminary review of secondary literature and the background knowledge of its creators. The latter constantly update it throughout the research process by adding new rows and populating existing columns with new information as they delve deeper into their data.

When inferring causality, *time* is of the essence. In historical institutionalism, time indicates the beginning, duration, and end of path dependencies and other causal chains,[13] and is a crucial indicator in marking critical junctures.[14] Process tracers equally seek to identify exact timeframes and attach events and interactions to specific points in time.[15] Therefore, it is imperative to introduce uniform date and time units as the first item in the codebook; every row in the spreadsheet could represent a month, a day, or even an hour or a minute, depending on the level of granularity of inference. The resulting timeline constitutes the first layer of research. Once the timeline is identified, documented, and verified, it provides a foundation for multi-layered analysis that may include temporal, spatial, and networked elements.

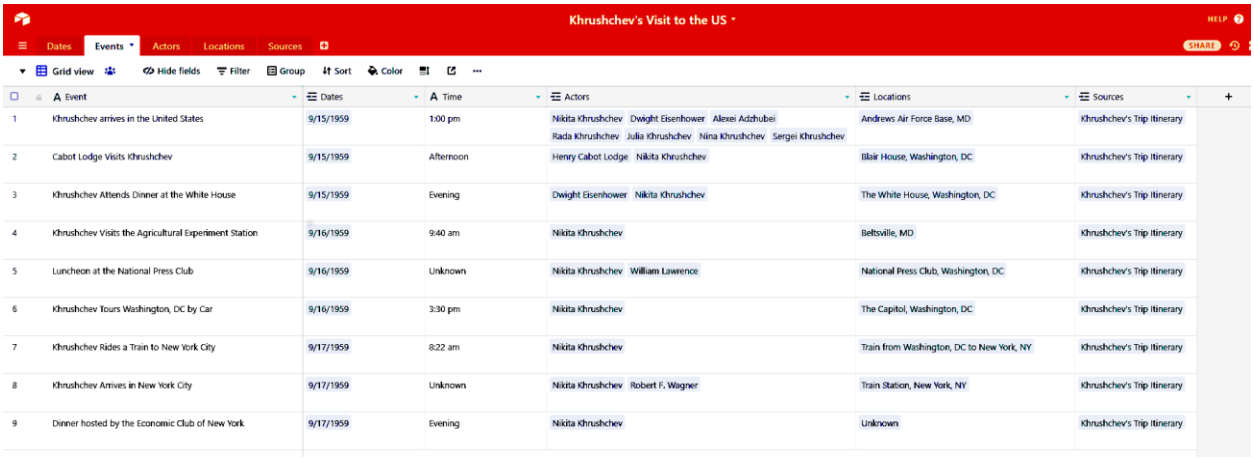
For the codebook to be truly effective, each of the four categories – *time*, *events*, *actors*, and *locations* – should have a separate sheet or tab. Each row in those sheets should contain, in separate columns, information regarding the three other categories, such that a row in the *actors* sheet that is dedicated to an individual would include all the *dates*, *locations*, and *events* that feature that individual under columns named as such. Thus, for example, in a hypothetical study of Soviet Premier Nikita Khrushchev's 1959 visit to the United States, one row in the *events*

[13] Pierson, *Politics in Time*, especially 44–46.
[14] Capoccia and Kelemen, "The Study of Critical Junctures," 348–51.
[15] Collier, "Understanding Process Tracing," 824; Ricks and Liu, "Process-Tracing Research Designs," 2.

sheet may be dedicated to the *event* "Khrushchev's visit to Disneyland canceled." In this row, the cell under the *time* column would read "19 September 1959." It might even present the hour or the part of the day in which the event occurred. The cell under the *actors* column would contain the names "Nikita Khrushchev" as well as "LAPD Chief William Parker," who recommended canceling the visit. The cell under the *locations* column would read "Los Angeles, CA." If the codebook is designed as a relational database,[16] all these items would be interlinked, such that clicking on the "19 September 1959" cell would take users to the "19 September 1959" row in the *time* sheet, where they would find other *events* (for example, "Khrushchev Attends a Reception at the Ambassador Hotel"), *actors* ("Marilyn Monroe" or "Frank Sinatra"), and *locations* ("Los Angeles-San Francisco Train") featured on that date. Figure 3 demonstrates how the *events* tab might look in a database depicting the first day of Khrushchev's visit.

[16] On relational databases and their creation see Michael J. Hernandez, *Database Design for Mere Mortals: A Hands-on Guide to Relational Database Design*, 3rd ed. (Upper Saddle River, NJ: Addison-Wesley, 2013).

FIGURE 4: AN EXAMPLE OF THE *EVENTS* SHEETS IN AN UNSTRUCTURED BIG DATA CODEBOOK



Source: A screenshot taken from the spreadsheet/database service Airtable *(https://airtable.com) tracing the first day of Khrushchev's visit to the United States. Data are from the web page "Khrushchev's Trip Itinerary" on the website of the PBS 2014 program* Cold War Roadshow *(http://www.pbs.org/wgbh/americanexperience/features/cold-war-roadshow-nikita-khrushchevs-trip-itinerary). For a copy of this example database in Excel format, see replication materials at https://github.com/for-anonymous-review/Big-Data-in-Political-Science.*

To further improve and contextualize the retrieval of relevant details, researchers can allocate additional columns to *descriptive metadata* – data that describe the content of a resource and allow users to look up that resource based on its attributes.[17] Such metadata can include researcher-generated keywords related to the item, such as "Soviet Union" and "leader" in the case of the *actor* Nikita Khrushchev or "disarmament" and "Berlin" in the case of the *event* "Khrushchev Visits Camp David." Since even the simplest spreadsheet applications offer the

[17] Richard Gartner, *Metadata: Shaping Knowledge from Antiquity to the Semantic Web* (Cham, Switzerland: Springer, 2016), 6–7; Richard Pearce-Moses, *A Glossary of Archival and Records Terminology*, Archival Fundamentals Series (Chicago, IL: Society of American Archivists, 2005), 113.

ability to search, sort, and filter text strings (groups of characters), such tagging would facilitate the retrieval of observations (rows) and data sources. Other columns can include the definition or description of the item in question or the researchers' comments and insights.

By creating such a codebook, political scientists can organize the immense pool of dates, names, and places found in unstructured big data in a logical manner without losing any connection to their sources. The structured nature of the database would allow researchers to establish temporal sequences, trace processes, test hypotheses, and make causal claims. Instant access to every relevant piece of raw data would enable close reading, offset the loss of context that is characteristic of structured big data, and satisfy the "burden of proof" demand.[18] By documenting every decision when collecting the data and organizing them in a database, researchers can guarantee that their progress is replicable and, if needed, reversible. Moreover, while we focus here on the applications of these practices to within-case research designs, such fine-grained and time-sensitive data can also serve as a basis for the creation of more reliable structured datasets for quantitative analyses featuring a large number of cases, given that data in the codebook are already structured to a great degree.

One caveat of this method is that it focuses on research that relies heavily on detailed unstructured evidence and that seeks to establish temporal sequences for making causal claims. This approach to analyzing unstructured data corresponds with the principles of historical institutionalism and process tracing as discussed in the article; scholars working in both traditions underscore the significance of nuanced and hypothesis-driven research for establishing causality. That is not to say that we do not embrace other research methods, including purely

[18] On the "burden of proof" see Lubet, *Interrogating Ethnography*, 3–4.

quantitative ones, for dealing with unstructured data. We believe that scholars should choose their methods in accordance with the specific objectives of their research project.