



## NATIONAL ENDOWMENT FOR THE HUMANITIES

DIVISION OF RESEARCH PROGRAMS

### **Parts of a Successful Application**

The attached document contains the narrative portion of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and applicants are urged to prepare a proposal that reflects their unique project and aspirations.

Prospective applicants should consult the application guidelines at <https://www.neh.gov/grants/research/neh-mellon-fellowships-digital-publication> for instructions.

Applicants are also strongly encouraged to consult with the NEH Research Division staff well before a grant deadline.

This attachment only contains the narrative, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

**The application format has been changed since this application was submitted.** You must follow the guidelines in the currently posted Notice of Funding Opportunity (see above link).

**Project Title:** *A computational analysis of group representation at U.S. Congressional hearings since 1877*

**Institution:** University of California Boulder

**Project Director:** Vilja Hulden

**Grant Program:** NEH-Mellon Fellowships for Digital Publication

# A computational analysis of group representation at U.S. Congressional hearings since 1877

## SIGNIFICANCE AND CONTRIBUTION

Since the early twentieth century, the United States Congress has in most years held hundreds and even thousands of hearings to investigate various societal problems and collect viewpoints on proposed legislation. Who has had the ear of the national legislature as it has deliberated on laws to govern our common public life? Who gets to speak to the state—and who does the state listen to?

This project aims to understand long-term patterns in representation at Congressional hearings. It does so by conducting a large-scale computational analysis of hearings over the past 140 years using available metadata—information like hearings title, committee, subjects, and witnesses and witness affiliations—and the full text of hearings. Analyzing metadata can answer questions like which groups have been prominently represented and how that has changed over time, or which committees have been key in considering particular topics, or what kinds of witnesses dominate at what kinds of hearings. For example, how has the representation of labor vs. business changed with political shifts or the prominence of unions? What topics get the most testimony from academics? At what committees do women’s groups mostly testify? Text mining and analyzing the full text will complement such metadata analyses by illuminating how the topics of hearings have fluctuated over time and how language used in testimony about particular topics or by particular groups varies: when has the Congress been particularly concerned with topics related to, say, motherhood, or how has the language witnesses use about the environment changed over time? Both types of analysis will be contextualized using “traditional” scholarly methods like reading witness statements and selected archival materials and making use of existing scholarship.

As the most accessible form of (federal) lobbying, Congressional hearings offer a unique window into the functioning of American democracy. Indeed, increasing the number of Congressional hearings was once advocated as a cure for secretive, money-driven lobbying. Even if the remedy proved insufficient, Congressional hearings nevertheless provide a public way for citizens to speak to their representatives and for representatives to gather viewpoints on a particular topic of legislative interest. Yet, despite the use of Congressional hearings as source material on a variety of topics and some work on the internal dynamics of specific committees (see e.g. Greene, 1998; Witwer, 2003; Auerbach, 1966; Dragna, 1976; Olzak and Soule, 2009; Gelb and Palley, 1979), there is a dearth of scholarship on Congressional hearings as a research topic in their own right, especially for earlier historical periods. The vast majority of work related to interest groups and hearings is focused on the period after about 1960 (see e.g. Albert, 2013; Burstein, 2014; Soule et al., 1999; Olzak et al., 2016; Kollman, 1997), and unlike this project, generally targets a fairly short time span. Indeed, on the whole, there is less work than one might expect on historical lobbying. The main work on historical patterns in testimony before Congress is by Daniel Tichenor and Richard Harris, who, pointing out that “political scientists know precious little about the contours of interest group politics in the United States before the 1960s,” use Congressional hearings metadata to map out a broad overview of the rise and fall of interest groups for the period 1833–1917 (Tichenor and Harris, 2003; Harris and Tichenor, 2009; Tichenor and Harris, 2005, 253).<sup>1</sup> This project aims to move beyond the coarse-grained and in their own words preliminary analysis offered by Tichenor and Harris to examining hearings metadata as well as testimony content in the context of a number of historical debates.

In addition to contributing to a better historical knowledge of representation at Congressional hearings, the project also engages scholarship in specific subfields of history. Preliminary work I have completed on the representation of labor and business groups at historical Congressional hearings suggests that labor’s strength at hearings has fluctuated in response to both labor’s electoral alliances and to its strength in the

---

<sup>1</sup>Of course, there are several works that touch on particular aspects of historical lobbying: for example, an old classic by E. Pendleton Herring examines the 1920s rise of interest groups, though he is not particularly focused on Congressional hearings (Herring, 1929), while Elisabeth S. Clemens investigates historical lobbying by grassroots organizations (Clemens, 1997).

broader civil society, and that the concerns of labor representatives have stretched well beyond bread-and-butter issues to social welfare and civil rights. Such findings have clear relevance to debates within labor history over whether a labor union is fundamentally a social movement or an institution, whether workers are best served by “responsible” negotiation strategies and political alliances aiming to build long-term institutional presence or by radical protest and critique from the outside, whether the state is better viewed as a repressive enemy or a protective friend to labor, and whether American “business unionism” has done too little to engage the state on broader social issues (see e.g. Moody, 1988; Burns, 2010; Lichtenstein, 1982; Robertson, 2000; Feurer and Pearson, 2017). They also underline the need to see the state and the legislature as embedded in, not separate from, the wider society.

The final product of this project will be a multifaceted digital work (using Scalar, a popular platform for digital publication of academic work), intended for publication in the Stanford University Press Digital Publications series, where the project has the support of the senior editor, Friederike Sophie Sundaram, and where it is now under consideration for future peer review. The digital work is aimed at scholarly and general audiences in equal measure. It constructs a narrative that engages scholarly arguments about lobbying and representation, but it also offers a modular structure and multiple pathways that invite students and the general public to explore who and what Congress has paid attention to over the years. The digital medium is crucial not only in creating interactive data visualizations, but also in demystifying the construction of the arguments from data. The format rhymes with the project’s intellectual theme of democracy and representation by inviting the reader to participate and highlighting rather than hiding the extent to which choices about who and what issues are present on the stage shapes the argument.

## ORGANIZATION, CONCEPTS, AND METHODS

This project treats Congressional hearings as a research object, employing computational data and text analysis as well as close (i.e., traditional) reading of the material. Part of the project relies exclusively on metadata—information like the hearing title, date, committee and subcommittee, witness name, or subjects—that can be processed to enable one to examine representation by geographical location, group affiliation, per committee, and so on. Another part relies on the full text of the hearings. While the available full text has some limitations,<sup>2</sup> methods that rely on a bag-of-words approach (where the order of words in a text does not matter) as well as methods that rely on word proximity within a particular window (i.e., methods based on context) can be deployed productively. The project will primarily make use of topic modeling (the automated extraction of a set number of topics based on word cooccurrence within documents) and word embeddings (somewhat simplified, transforming words into numerical representations based on the context words with which they appear, enabling calculations of similarity between words.) Topic modeling can be used to e.g. examine how the topics of testimony by a particular group (labor, business, women’s groups etc.) shift over time, or how testimony by different groups at the same hearings differs in topics. Word embeddings can do something similar for vocabulary and word meanings: say, how the vocabulary regarding the environment has changed over time, or whether labor representatives and business groups deploy words like *economy* or *jobs* in different senses.

To make full use of the possibilities of large-scale data while remaining sensitive to historical context, the project examines Congressional hearings from a birds-eye view as well as through in-depth case studies. The birds-eye perspective draws on the possibilities of big data to provide overviews and to introduce Congressional hearings as well as the project’s methodology to the reader unfamiliar with either the procedures of Congressional committees or metadata and text analysis. The case studies ensure that the data analysis is firmly embedded in a specific context and interpretive framework and engages with historical content and existing scholarship—that it is not limited to questions of *what*, but can also help us understand *how* or *why*.

---

<sup>2</sup>For example, some “stopwords” like *have*, *and*, *that* have been removed, making full syntactic parsing impossible.

The case studies will illuminate differences in representation: who is (this form of) American democracy open to, and what does it take for “underdogs” to get a hearing before Congress? At present, I envision three case studies. One draws on my expertise as a labor historian to compare labor and business representation at hearings, combining the hearings metadata with other data sets (union density, strike frequency) to examine labor’s shifting fortunes at hearings. Another focuses on women and women’s groups, aiming to understand “women’s issues” before Congress both before and after suffrage. A third case study is organized around a topic rather than a witness group, examining hearings related to environmental issues. It zeroes in on the 91st and 92nd Congresses (1969–1973) when the environment emerged somewhat suddenly as a new focus of public policy, but places those discussions within a longer context of how language about the environment has shifted over the course of the twentieth century. All the case studies track shifts in types of witnesses, witness group concerns, and language used in testimony to build a holistic and long-term view of who is represented at Congressional hearings and how their concerns are talked about.

The case studies link to questions of fundamental relevance to American democracy. Work and its regulation engage core cultural values about self-reliance, an “American” standard of living, fairness, and the freedom of assembly; testimony by women’s organizations highlights issues of democratic inclusion and exclusion based on group characteristics; the question of environmental protections raises basic questions about the reach of politics and governmental regulation and the difficulties of securing the common good.

## WORK PLAN

Much of the effort so far has focused on obtaining and processing the data. That task is now mostly completed, and further processing can focus on improving the categorization of witnesses. I have presented preliminary analyses of the data at various conferences (American Historical Association, Organization of American Historians, Labor and Working Class History Association) as well as published a short analysis in volume 2 of *Current Research in Digital History*. The feedback at conferences and from anonymous readers has been invaluable in molding the project into its present shape, with equal emphasis on engaging with scholarly arguments and on engaging a wider readership in thinking about both scholarly (and digital) methodology and the shape of our democracy.

The funding sought here would enable me to devote full-time attention to the project at a crucial stage. The data has been acquired and preliminarily processed and sufficient preliminary, exploratory research has been carried out to make productive work feasible. At the same time, the research has reached the point where exploratory forays into the data are no longer sufficient, and the project would benefit greatly from an intensive and immersive research stretch to complete one of the case studies fully and to invest in conceptualizing and sketching out the form of the final product. At the end of the funding period, I aim to be ready to submit the first case study for peer review.

*January 2021:* Resolve remaining issues with data processing.

*February–March 2021:* Complete identification of organizations in metadata and finalize creation of multiple levels of broader categorizations to enable flexible future analyses. Separate and label full-text testimonies by these categories (using metadata.) Carry out statistical and visual analyses of representation data extracted from this metadata with other sources of data (union density, strikes, etc., nationally and by industry/state.)

*March–May 2021:* Begin intensive processing of labor witnesses at hearings, including full text. Compare bodies of testimonies by category affiliation in terms of topics (from topic modeling) and word meaning (as measured by word vectors.) Do labor’s witness statements combine topics differently than those by corporate representatives? Are there key words that each uses to mean very different things (as determined by context)?

*June 2021:* Take stock. What are the most significant or suggestive findings so far? Go back to the literature and (re)read with the findings in mind.

*July 2021:* Rethink analyses and create new ones as necessary.

*August–December 2021:* Writing and creation of visualizations. The goal is to write two chapter-length analyses (one chapter on overall labor vs. business representation since 1877, aiming to explain how and why it has shifted, and another delving into a specific legislative issue to enable a closer and more contextualized examination of the intention and impact of labor testimony) as well as to create visualizations that help illuminate those analyses, but also invite exploration and speculation.

*Concurrent and beyond funding period end:* Experiment with different visualization tools and code libraries; pursue opportunities to do collaborative work with CU Boulder visualization and digital humanities classes; use labor analysis as a guideline to shape the further two case studies; complete final product.

#### COMPETENCIES, SKILLS, AND ACCESS

The data for the project comes from the ProQuest Congressional database that contains metadata and full text for Congressional hearings since 1823; this project focuses on the period since 1877, before which Congressional hearings were far more infrequent (Sevetson, 2011). The XML-format data was acquired for me by my university library; I have subsequently written Python scripts to process and clean most of the metadata into a CSV (spreadsheet) format, extract the full text, and experiment with analyses of both.

I have experience with analyzing spreadsheet data as well as with using topic modeling and word embeddings analysis, and have published and presented conference papers drawing on these methods. I also possess the necessary skills to create the digital publication. I am well versed in HTML, regularly incorporate creating WordPress projects into my classes, and have created multiple web sites, some with SQL integration; I am also familiar with tools like Tableau and ArcGIS and can program in Python, R, and JavaScript. I have also begun to construct the digital framework for the current project in Scalar, which I find quite user-friendly. My university also has good support for digital scholarship and I have collaborated extensively with our highly competent digital scholarship librarians and research data experts.

#### FINAL PRODUCT AND DISSEMINATION

The final product will be a digital “book” constructed using Scalar. It is projected to be published in the Stanford University Press Digital Publications series; we have preliminarily agreed on a first round of peer review after the completion of the first case study.

As a more flexible narrative format than a print book, the digital work allows multiple, non-linear pathways embedded in a larger narrative frame, thus emphasizing the multiplicity of contexts among which it may not be necessary to choose one as the privileged, determinative one. The ability to display the results of data analysis in interactive as well as static formats enables the work to make explicit the way in which choices of emphasis shape the argument, and allows the work to guide the reader in extracting different meanings from the data. A fundamental goal of the project is to underline the fact that choices and interpretations are always part of every stage of the research process, a point that echoes with the intellectual focus of the project on how representation in the democratic process has shifted over time and shaped our society.

My hope is that the modularity and interactivity embedded in the work will make it more useful for the general public and for classroom use. A modular structure and a variety of shorter and longer texts allows the general public to select the portions of the work they wish to engage with. Insight into the methodology should also prove useful for classroom use both in courses focusing on humanities data and in traditional history courses looking to illuminate how historical sources and data become narrative arguments. Many teachers and professors (myself included!) struggle to illuminate to students how scholarship is made, how interpretation necessarily shapes the meaning of the data and the sources rather than being just “bias,” and how arguments and different types of evidence relate to each other. The final product hopes to provide a transparent view of the gears and pulleys—and kludges—behind the narrative, rather than hiding the messiness of scholarship.