Diplomatic Documents Data for International Relations: The *Freedom of Information Archive* (FOIArchive) Database[1]

Matthew J. Connelly[2]   Raymond Hicks[3]   Robert Jervis[4]

Arthur Spirling[5]   Clara H. Suong[6]

7495 words

Abstract

We introduce the *Freedom of Information Archive* (FOIArchive) Database, a collection of over 3 million documents about state diplomacy. Substantively, our database focusses on the US and provides opportunities to analyze previously classified (or publicly unavailable) corpora of internal government documents which include raw---often full---text of those documents. We also provide within-country diplomatic records for the US, UK, and Brazil. The full span of the data is 1620-2013, but is mainly from the 20th century. Our database allows scholars to view text and associated statistics online, and to download and view customized datasets via an API. We provide extensive metadata about the documents, including the countries and persons they mention, and their topics and classification levels. The metadata includes information we extracted with domain-specific, customized Natural Language Processing tools. To demonstrate the potential of this data, we use it to design and validate a new index for "country importance" in the context of US foreign policy priorities.

Keywords

Diplomacy, diplomatic communication, US foreign policy, text-as-data, historical documents

[2] Professor of History, Columbia University, mjc96@columbia.edu
[3] Project Manager, History Lab, Columbia University, rh2883@columbia.edu
[4] Professor of International and Public Affairs, Columbia University, rlj1@columbia.edu
[5] Associate Professor of Politics and Data Science, New York University, as9934@nyu.edu
[6] Postdoctoral Associate, Department of Political Science, Duke University, clara.suong@duke.edu

1. Introduction

International Relations is by definition the study of states' interactions with other states. Since at least the time of the Ancient Egyptians, these relationships have been directed, promoted, and negotiated by way of diplomacy and diplomats. Unsurprisingly then, scholarship has devoted considerable attention to the subject of envoys and their activities. As with all research efforts, data is key to progress. But data on diplomacy, whether qualitative or quantitative, can be difficult to obtain, and difficult to work with. This is true even for wealthy, democratic nations with aggressive freedom of information provisions such as the United States.

The reasons for this paucity are well rehearsed. For one thing, diplomatic records are often covered by state secrecy laws, and may be made available for public consumption on a haphazard schedule, if at all. Simultaneously though, governments often release files in overwhelming numbers but without careful curation, making it impossible for scholars to keep up with the flow of information available to them. On the one hand, its sheer size makes it difficult to catalog and work with, especially if one wishes to move between aggregate analysis and the inspection of individual cases. On the other hand, finding more than one example of a particular phenomenon and applying the scientific method to it can be daunting. All told, there is an obvious need for a database of diplomatic data, for the US and beyond.

For this reason, this paper introduces a publicly available database with over three million previously unreleased government documents, focused on the US but including records for the UK and Brazil too. The focus of our database is on comprehensive collections of diplomatic documents, especially in the born-digital age. We also introduce an online platform with an application programming interface (API) and website with which scholars can search for, read through, and download datasets customized to their research needs.

This database, *Freedom of Information Archive* (FOIArchive), supports the text-as-data approach to research on International Relations (IR). It allows scholars to study multiple aspects of diplomacy, but especially intra-state relationships: that is, diplomatic priorities, organization and communication *within* states. In addition to the full text of documents, scholars can access many different types of metadata, including entities and topics identified using natural language processing and machine learning methods.

We will describe our collection in some detail shortly. First, we discuss prior research on diplomacy and our motivation for creating the database and platform.

2. Motivation: Inter-State and Intra-State Themes in IR

The study of diplomacy in International Relations has given rise to a large literature with inter-state and intra-state themes. Our data speaks to both, but primarily to the latter.

The *inter-state* diplomacy literature has several themes, with one dividing principle being the approach to adversaries in the international system. While scholars have emphasized the centrality of accumulating "power", some nonetheless recognize that compromising with other states is a common way to do this (e.g. Morgenthau 1948). By contrast, a more recent formal literature on coercive bargaining (e.g. Fearon 1994) considers the importance of adversarial diplomacy on disputes. This includes work on diplomatic communication *per se* (e.g. Kurizaki 2007; Trager 2017). In this literature, credible communication and reputations (Sartori 2002) are key. On the empirical side, the focus is measuring resolve from diplomatic statements (e.g. Katagiri and Min 2019) which may be in part obtained from private documents of governments (Gill and Spirling 2015). The communication literature is predicated on states' beliefs and their effect on international political outcomes. How diplomacy affects these beliefs remains important in the political science literature (e.g. Jervis 1976).

The procedural activities and results of diplomatic interactions have also been of considerable interest. For instance, scholars have investigated diplomatic representation (e.g. Singer and Small 1966), explored the formation of diplomatic networks (e.g. Neumayer 2008), and analyzed the effect of bilateral diplomatic ties on formation of preferential trade agreements (Plouffe and van der Sterren 2016).

As regards *intra-state* relationships, the IR literature on diplomacy has shifted from analysis at the macro-level, such as diplomacy between powers (3rd image), to the micro-level (1st and 2nd images) and an effort to understand how diplomacy "works" within states. In addition to studying historical cases, scholars have built their theories about diplomacy from the "ground-up", sometimes testing them on individuals in experimental settings (e.g. Hall and Yarhi-Milo 2012; Holmes 2018; Wheeler 2018). They have also analyzed the causes and consequences of diplomatic appointments and meetings (e.g. Gertz 2018; Haglund 2015; Gray and Potter forthcoming).

The growing body of research on diplomacy has greatly advanced the study of international relations. But methodological problems remain. For one, IR scholars find it difficult to obtain data that would best test their theories: especially those between agents of the same state. Records that reveal the private beliefs and communications of elites are not immediately available, and even when declassified, typically require a close reading of individual records. This calls for a multi-disciplinary research community, which would combine quantitative analysis of datasets and archival research in documents (Gerring 2012; Sagan 2014; Trachtenberg 2006). But this already formidable challenge has grown as governments have reduced researcher support in traditional archives. Instead, they are making millions of electronic records available through web-based search engines, but without providing any easy way to download the underlying data for quantitative analysis. These problems motivate a solution we provide below: new, curated, easily-extended, machine-readable collections that allow researchers to combine quantitative and qualitative approaches.

3. Contribution: What the FOIArchive Database Provides

International relations scholars have recently turned their attention to providing and analyzing text data sets, though these have been focused on political speeches (see e.g. Baturo et al 2017). To our knowledge, FOIArchive is the first corpus to bring together large scale numbers of machine-readable, intra-state communication documents for analysis. Our collections include data on both private (i.e. classified) and public information and actions, diplomacy at multiple levels of analysis, and communication across diverse topics. This will be especially helpful for those studying decision-making within states. New data will be added with each release of newly-declassified documents.

We will describe the documents in detail below, but here we highlight two technical innovations in our efforts. First, we have undertaken Named Entity Recognition (NER) of persons, places and organizations automatically detected in, and extracted from, the diplomatic documents using customized Natural Language Processing tools. Our technical efforts here improve over "off-the-shelf" products: we have developed a standalone specialized system that extracts titles, pronouns, and country adjectives (such as Israeli or French) that conventional implementations miss. We also distinguish between types of Geographical/Social/Political Entities (GPE)[7], separating out mentions of countries as governments from countries as locations.[8] Beyond their intrinsic usefulness, we hope our NER efforts will inspire follow-up computational analyses of historical documents in IR research.

In addition to the extracted information about Named Entities, our data includes information about the thematic diversity in diplomatic communication. We use the canonical Latent Dirichlet Allocation (LDA) "topic" model to do this (Blei, Ng, and Jordan 2003). The topics are meant to give a quick overview of the content of a document and to allow users to easily find similar documents. We wrote a Python script to tokenize the data and run the topic model with the same parameters and number of topics across collections. The Python script could be easily adapted to allow a different number of topics. While the techniques can be powerful, labeling output from such models (i.e., understanding the themes the topics represent) is not trivial: there may be many "junk" topics, or the topics may be substantively interesting but require detailed inspection to interpret. We therefore make two versions of the topic modeling data available; both versions will be available as metadata and on the website. In one version, historians have curated the results, reading through the documents in a topic to make sure that at least half the documents assigned to a topic fit the topic, and discarding the other topics. In the other, we provide all the data in uncurated form, with the top tokens rather than labels for all the topics. With the uncurated topic data, researchers will have to read through a selection of documents with the same topic themselves to determine the underlying label. A latent topic label cannot be adequately determined from the list of tokens only so we discourage researchers from reading too much into the topic tokens without taking a deeper look at the documents. For some collections, this

---

[7] "Composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.)" (Linguistic Data Consortium 2005, 13)
[8] We are still working on improving this part of the parser to reliably distinguish the two.

uncurated version will also be more complete, since it can include the most recently released documents[9]

4. The FOIArchive Database

In this section, we describe the selection criteria for the eight groups of government documents (corpora or "collections") in the FOIArchive Database, and the strengths and limitations of the data that is currently available. We then discuss the process of collecting, processing, and integrating the documents in our database before introducing the online platform from which the data will be publicly available. This is followed by in-depth discussions of our two largest collections, the State Department Central Foreign Policy Files and the Foreign Relations of the United States series.

4.1. Data Overview

FOIArchive was designed for the era of "born-digital" electronic records, which in the U.S. began with the State Department *Central Foreign Policy Files* (CFPF) in 1973. This is the largest part of the FOIArchive in terms of the number of documents, and it grows each time the National Archives releases another yearly installment. To understand the centrality of the CFPF, note that the National Archives recommends that "research on almost all topics relating to U.S. foreign policy and foreign affairs should begin with the central files of the Department of State and then branch out to other records of the Department of State and to the records of other foreign affairs agencies."[10]

Going forward, the goal is to add "comprehensive" and official collections of diplomatic records, especially born-digital documents. By comprehensive, we mean that we will collect all released records of a collection rather than a subset. We have expanded the coverage of the post-1973 era by adding digitized collections that include these years but extend the chronological range (e.g. the *Foreign Relations of the United States* series, *Declassified Documents Online,* the *President's Daily Briefs,* and the *Henry Kissinger Telephone Conversation Collection).* We have prioritized the creation of a critical mass of U.S. documents, but have also added foreign collections that make new kinds of research possible, such as the *Cabinet Papers* from the UK, and the *Azeredo da Silveira Papers* from Brazil. Finally, we seek to include new kinds of electronic data that will be increasingly important in the future, e.g. the *Hillary Clinton Email Collection.*

Our *Central Foreign Policy File* collection includes full text and associated metadata of the "cables" exchanged between and among US diplomatic posts and the State Department's headquarters in Washington, DC from 1973 to 1979. The documents exclude any communications between the State Department and other government agencies. Our database makes available metadata and full-text from 2,081,276 State Department cables and the metadata from 1,133,017 other records, including cables with still-classified message text,

---

airgrams, and other paper records.[11] This corpus was parsed from all the source XML (Extensible Markup Language) files currently available from the US National Archives. Because the collection contains communications between embassies and the State Department on a variety of issues, it is very useful for understanding the day-to-day activity of the State Department, including diplomats' role in trade and foreign investment policy and their understanding of the political events in a country.

Our data also includes 307,533 documents from the *Foreign Relations of the United States* (FRUS) series and their metadata, a published collection chosen by State Department historians as the most important records from across the federal government. The collection spans the period of 1620--1989. This corpus was also parsed from XML files downloaded from the State Department's GitHub site. These documents are more selective and shed light on responses to the most important foreign policy priorities in different geographic and issue areas.

President's Daily Briefs (PDBs) are a collection of daily reports to the President, Vice President, and select officials summarizing the most important information and analysis from the intelligence community, including the CIA and the NSA, from the Kennedy, Johnson, Nixon, and Ford administrations. The collection also includes the *President's Intelligence Checklists* (PICLs) from the Kennedy administration. Our database makes available 5,011 declassified PICLs and PDBs from 1961 to 1977. The source documents were "scraped," or copied, as PDF (Portable Document Format) files from the CIA's online Freedom of Information Archive (FOIA) reading room and parsed. While a lot of information from the PDBs is still redacted, the text can be used as an indicator of the important intelligence issues and areas facing the president over a number of years.

In addition, our data includes 4,552 transcribed phone calls from Henry Kissinger's time as Secretary of State, 1973-1976, as well as 54,149 emails that were sent or received by Hillary Clinton while she was Secretary of State from 2009 to 2013. Both corpora were acquired by scraping and processing the PDF files uploaded on the State Department's FOIA reading room and both offer candid insights into the day-to-day practice of political diplomacy.

Our database also stores metadata of the 117,509 documents in the *US Declassified Documents Online* (DDO) collection, spanning the years 1900-2008, but especially covering the Cold War era. The collection mainly consists of documents from Presidential libraries that researchers requested to have declassified, which were then provided to Gale-Cengage. They are identified by the originating agency or department, including the Defense Department, the FBI, the State Department, and other agencies. We parsed the source XML files and images provided by Gale Cengage to create our collection.

---

[11] The text for most of these non-cable records are in the P-Reel collection and are available at the National Archives, but only in paper format (603,362). About half were "withdrawn" during the declassification process (529,655).

Additionally, our database provides the content of the UK *Cabinet Papers* and the *Azeredo da Silveira Papers*, permitting researchers to investigate both sides of bilateral diplomacy. The *Cabinet Papers*, spanning 1907-1990, include Cabinet Conclusions, which are essentially minutes of cabinet meetings; Cabinet Memoranda, which are the reports and papers for briefing cabinet ministers prior to cabinet meetings; Cabinet Secretary's notebooks, which are the Cabinet Secretary's handwritten notebooks that often constitute the first draft of the Conclusions; and the Cabinet Office precedent books, which describe the Cabinet Office and its procedures.

The *Azeredo da Silveira Papers* consists of personal and official documents from 1973 to 1979 collected by the Brazilian Minister of Foreign Affairs. The 10,279 digitized documents include letters, memos, and other correspondence, and are considered among the most important for documenting Brazilian diplomacy in this period. The documents were provided by the Center for Research and Documentation of the Contemporary History of Brazil (CPDOC) at Fundação Getulio Vargas (FGV), where da Silveira's private papers are housed. The collection connects with and complements other 1970s corpora in FOIArchive. It also includes documents in multiple languages, a test-bed for research in which language, e.g. translations, is of interest.

Altogether, FOIArchive currently includes 3,657,378 documents. Of these, we provide full text of over 2.9 million documents.[12] Our data also covers a long time span -- the oldest document dates from 1620, and the newest is from 2013. But it is important to note the limitations. The collection is richest for the 1970s, the beginning of the era of electronic records, with multiple overlapping collections, and one that provides a near day-to-day inventory of all important activity of the U.S. State Department. But even the CFPF does not contain every cable---only those preserved by archivists. Far fewer records are available after 1979 and, with special exceptions like the Clinton e-mails, more can be added only when they are scheduled for declassification in years to come. For earlier periods, research possibilities are even more dependent on the scope and nature of individual collections, such as the selective nature of FRUS, and the way its content has changed over more than 150 years. Some collections, like Declassified Documents Online, derive from the declassification requests of individual researchers, and all collections are limited by the vagaries of even "automatic declassification."

One of the important advantages of making multiple overlapping collections available is to help researchers analyze these differences, and determine how different choices of data might lead to

---

[12] Note that the National Archives classify the documents in the CFPF collection available online into the following categories: Electronic Telegrams; Electronic Telegram Withdrawal Cards; P-Reel Document Index Entries; and P-Reel Document Index Entries Withdrawal Cards. P-Reel documents are documents available only in the P-Reel ("P" for paper) microfilm format and digital withdrawal cards are those "created by both the Department of State and NARA for classified or otherwise restricted telegrams and index citation entries to microfilmed records where the citations themselves contain classified or otherwise restricted information" (US National Archives and Records Administration 2019).

different conclusions. They can also identify and measure potential bias resulting from document withdrawals and redactions, such as by comparing the metadata of full-text and withdrawn cables. Researchers should be careful about combining documents from different collections, think about potential biases, and match the FOIArchive data to their specific research questions.

In Table 1, we show key attributes of our corpora by collection.

TABLE 1: Overview of the Corpora

Country	Collection/ Corpus	Date Range	Number of Documents	Number of Documents with Metadata and Full Text	Number of Documents with Only Metadata	Unprocessed Document Format and Source
US	State Department Central Foreign Policy Files (CFPF)	Jan. 1, 1973-- Dec. 31, 1979	3,214,293	2,081,276	1,133,017	XML, US National Archives
	Foreign Relations of the US (FRUS)	Nov. 3, 1620-- Jan. 2, 1989	307,533	307,533	0	XML, US Dept. of State
	President's Daily Briefs (PDBs)	June 17, 1961-- Jan. 20, 1977	5,011	5,011	0	PDF, CIA
	Henry Kissinger Telephone Transcripts	Jan. 2, 1973-- Dec. 24, 1976	4,552	4,552	0	PDF, US Dept. of State
	Hillary Clinton Emails	Mar. 9, 2009-- July 7, 2013	54,149	54,149	0	PDF, US Dept. of State
	Declassified Documents Online (DDO)	June 15, 1900-- May 12, 2008	117,509	0	117,509	XML, Gale Cengage
SUBTOTAL			3,703,047	2,452,521	1,250,526	
UK	Cabinet Papers	Oct. 19, 1907-- Dec. 13, 1990	42,539	42,539	0	PDF, UK National Archives
Brazil	Azeredo da Silveira Papers	Nov. 15, 1973-- Nov. 24, 1979	10,279	10,279	0	XML, FGV
TOTAL			3,755,865	2,505,339	1,250,526	

Note that some collections include documents with only metadata and no full text. In particular, for the CFPF, the full text is available only for the "Electronic Telegrams;" the full text of

"withdrawn" or "P-Reel" documents is unavailable. Metadata is available for all CFPF records. We also do not show the full documents of the DDO Collection on our site, but instead provide users a summary of the document (which is algorithmically derived using LexRank (Erkan and Radev, 2004)) and then link to the website of Gale-Cengage, for users who would like to see the original document.

4.2. Workflow

FOIArchive is composed of documents that we obtained from a source, processed, and ingested. In the pre-processing stage, we scraped the source documents from their websites as XML or PDF files, extracted text from the documents, and cleaned the extracted text. To extract text from PDF files, we used an optical character recognition (OCR) library in Python, and then regular expression patterns to detect and extract different elements within documents.

During the processing stage, the pre-processed data at the document level was moved to an SQL database with separate databases for each collection. We processed the body text of the documents with NER tools to detect and extract Named Entities mentioned in each document. We also used Latent Dirichlet Allocation-based topic models to detect the topics discussed in the text of the documents. The pre-processed data as well as the Named Entities and topics we extracted from the text were saved as tables in our MySQL database that also recorded the relationships among them, allowing users to quickly create, compile, and download data related to distributions of key variables by collection. In this stage, the input was the preprocessed data and the output was the structured data present in the database.

Finally, in the post-processing stage, we checked the results and made modifications. In particular, we "curated" the results from the unsupervised topic models. We also reconciled entity names and other metadata to correct for typos or other errors.[13]

After the three stages, our MySQL database stores cleaned text of the documents, including their body text and metadata, and variables created by our additional processing, such as their Named Entities and topics. All three stages use Python scripts so that they can easily be rerun when new data are added, allowing us to quickly and reliably update our database. (See Figure 1 for a graphical representation of the workflow.)

Figure 1: Workflow

---

[13] To ensure compatibility with other sources we keep the metadata with errors as well.
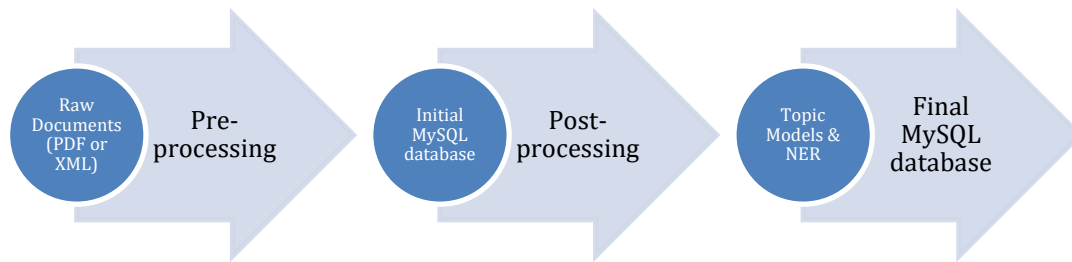
## 4.3. Online Platform: API and Website

Our database makes all of the documents and associated metadata accessible through an Application Programming Interface (API) and on the website. Researchers can use the API to download customized data in JSON format, including full-text of the documents. We have also created Stata and R interfaces with the API to allow users to download directly into either statistical package.[14]

The FOIArchive data is also viewable via a website (http://www.history-lab.org/) which is designed and maintained to make the same data accessible to a broader public through a graphical interface.[15] Figure 2 is a screenshot of the Document Explorer of the website to which users can configure and submit their queries and view the results:

---

[14] Draft versions of the Stata and R packages are available on request. We are currently updating and extending the API to add new features. The updates should be finished by Spring 2020. Code for both packages will be uploaded to our GitHub page with links to the packages also available on our website (history-lab.org). The website will also have a document with more detailed instructions for using the API.

[15] Earlier versions of the paper referred to the data as *Computational and Historical Resources on Nations and Organizations for the Social Sciences* or CHRONOS as it was titled in the NSF grant. There are at least 3 other NSF funded projects with CHRONOS in their name and an Italian company called Columbia makes a refrigerated display case called Chronos. To make our data easier to find we refer to it as its original name of Freedom of Information Archive (FOIArchive).

FIGURE 2: Screenshot of Our Website
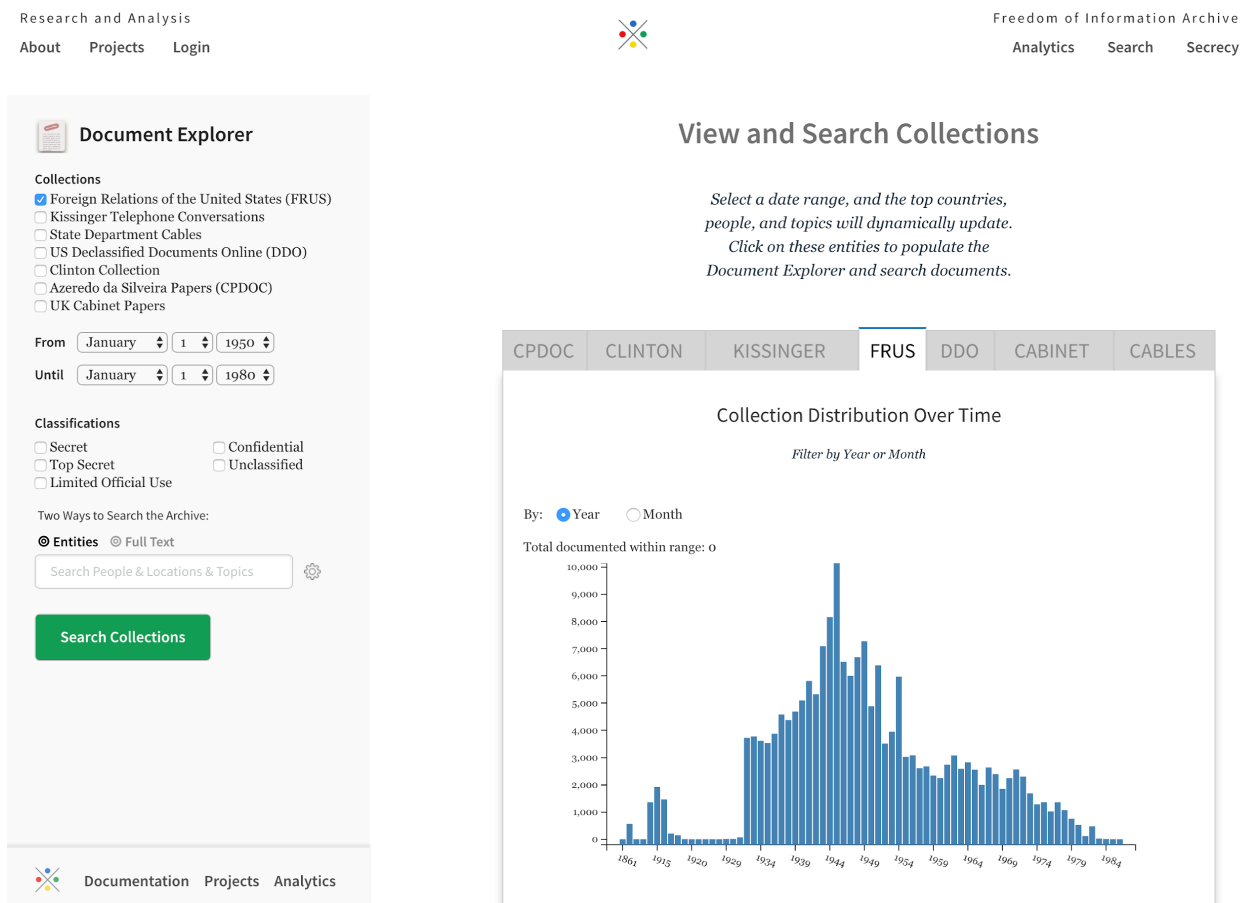
The interface includes tools to view collection directly, including the topic modelling data. Most importantly, once fully implemented, users will be able to download subsets of the data in a CSV format. For example, a scholar looking for all people mentioned in a given collection from 1968-1980 in documents highly relevant to a human rights topic in all collections will be able to download time-series data by specifying different filters.

4.4. Central Foreign Policy File Collection

The *Central Foreign Policy Files* (CFPF) collection is the biggest corpus in FOIArchive. The collection is comprised of communications exchanged between and among US diplomatic posts and the State Department headquarters in Washington, DC (sometimes referred to as "Main State" by diplomats after the headquarters' building). We include all those records that have undergone review for "automatic declassification," i.e. the years 1973 to 1979.

While the CFPF includes metadata from hundreds of thousands of records, such as airgrams and memoranda delivered by diplomatic pouch (what the National Archives calls "P-Reel" records, with P standing for paper), the most valuable part of the collection are the full-text diplomatic cables. They include reports on the political, economic, or security situation of the host state, often

including summaries of diplomats' meetings and interviews with local leaders or other sources of information. Cables also include queries or directives from Main State to diplomatic posts, such as orders to convey a certain message to the host government. Many more cables are mundane matters, such as organizing VIP visits or the internal management of diplomatic missions. But as with all archival collections, the State Department and the National Archives strive to identify and preserve only what they consider significant records. This appraisal process was done by subject (i.e. TAGS, to be discussed below). All records relating to political subjects were retained, but not those related to what were considered more routine or administrative matters, e.g. passports and visas, unless they were cross-referenced with other subjects deemed to be more substantive.

Despite an exponential increase in information gathered through other channels, such as foreign travels and the news media, officials in "sending" countries still rely to a great extent on confidential reports by their own diplomatic missions for information about another country (Kinne 2013, 248). In particular, officials in the sending country are extremely reliant on the embassies' "knowledge of the mind of the local leadership" (Berridge 2015, 123). For instance, the US government was dependent on reports from the US embassies in Cairo and Tel Aviv to "sens[e] the mood of Egyptian President Anwar Sadat" (Berridge 2015, 123) in the 1970s. This was "of vital importance to the Carter administration" when the US government was mediating between Egypt and Israel for a peace treaty that was eventually concluded in 1979 (Berridge 2015, 123).

Our data about this corpus include document-level variables that describe either each document's metadata or its content---the Named Entities and topics mentioned or discussed, in particular. Most, if not all, of the document-level variables can be aggregated into those at the corpus(collection)-level.

The following variables constitute key metadata of the documents in this corpus: id; subject; body; date; classification; from; to; tag; concept; office; and type. The variable "id" represents a unique identification number for each cable. The variables "subject" and "body" are for the cable's title and body text, respectively. The variables "from" and "to" display the US embassy or agency which transmitted a cable and the embassy or agency which received it. The "date" represents the relevant date for the cable, usually the date the cable was drafted.[16] The variable "classification" represents the sensitivity of the contents of the cable. Classification levels for the documents in this collection range from "Secret," "Confidential," "Limited Official Use," to "Unclassified" in the order of sensitivity. "Secret" documents are considered the most sensitive (causing the most damage if leaked). The National Archives has not yet made any "Top Secret" cables available in the CFPF collection, though some have been released through other channels. Similarly, the corpus may not include information that would only be available in intelligence community communications.

Below we display the number of cables in the CFPF corpus by month and classification level (FIGURE 3). The corpus includes over 3 million cables previously categorized as "Secret,"

---

[16] For some cables, such as those from 1978 and 1979 on the "P-Reel" that were withdrawn, the variable "date" represents the date on which a cable was sent.

"Confidential", "Limited Official Use" or "Unclassified." As noted above, political scientists often have theories about private information, and the plot demonstrates that our collections contain large numbers of documents able to speak to that interest (thousands of cases, for some years).

FIGURE 3: Number of Cables by Month
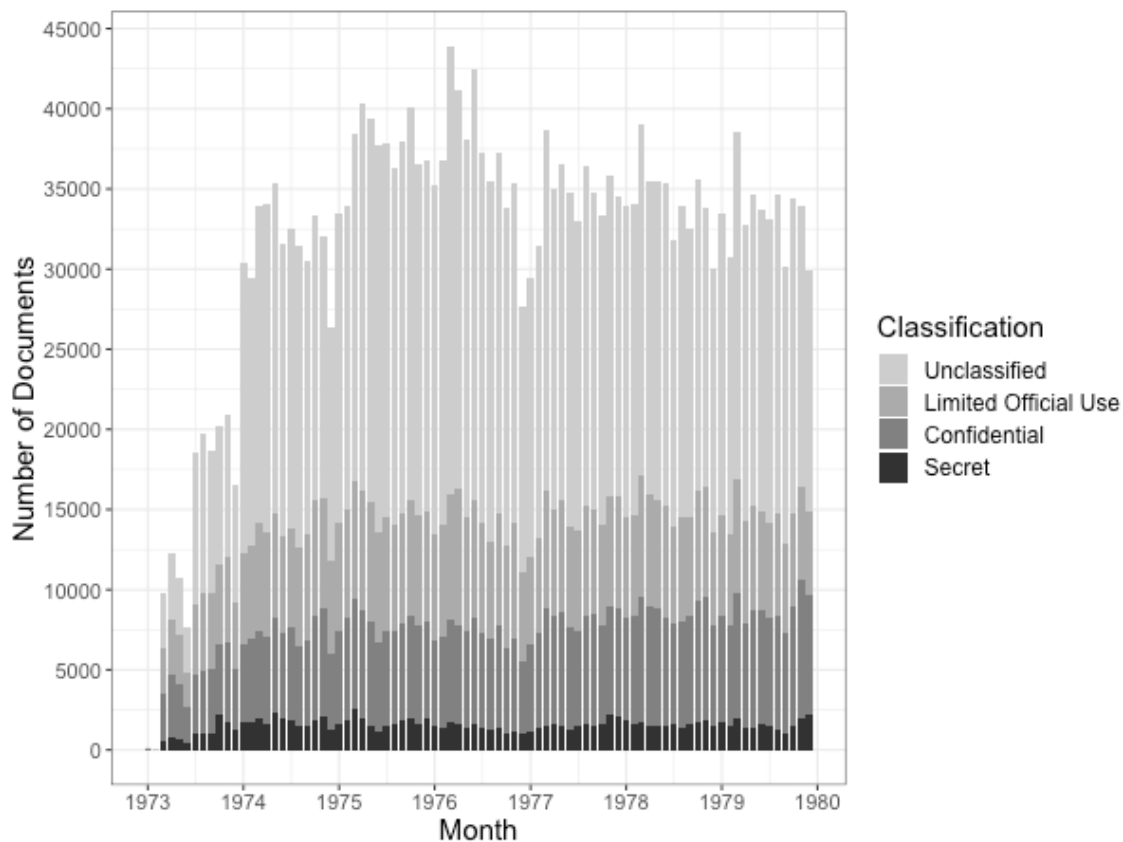
The variable "tag" represents the Traffic Analysis by Geography and Subject (TAGS) developed by the State Department for internal organizational and analytic purposes. It is a list of predefined fields (values) by the State Department that corresponds to a variety of subjects. Each cable will have some number of TAGS. These are then subdivided into particular areas of interest.

The variable "concepts" refers to the general concepts relevant to the content of the cables. Note that concepts are different from subject TAGS. For one, subject TAGS come from a predefined list. In contrast, concepts are open-ended, and it is up to the drafter of the cable to add whichever "concepts" they see fit to further refine the topic, subject, or nature of the cable. Concepts also tend to be abstract and broad in comparison to TAGS.

The variable "office" represents an internal designation for the specific office within the State Department or embassy which is to receive or which sent a particular cable. For example, the office designation "ORIGIN SS" identifies high-level communications that were routed through the Secretary of State's own office.

The variable "type" refers to the type of document. For instance, a cable with value "AI" for "type" is an airgram. "CC" stands for a "Congressional Correspondence," "DN" for a "Diplomatic Note," and so on.[17]

In addition, we created several variables by processing the body text of the cables. The variable "topic" refers to the topics discussed in the body text of each cable that were validated and curated by historians. We keep the top 3 topics for each document. The variables "person," "organization," and "gpe" refer to the Named Entities we have recognized, extracted, and classified from the text. "Person" represents each distinct person or a set of persons mentioned in each cable and can be classified into individuals, groups, or indefinite persons. "Organization" refers to each distinct organization or a set of organizations mentioned and can be government, commercial, educational, or non-governmental organizations. The variable "gpe" represents Geographical/Social/Political Entities (GPEs) extracted from the text.

Figure 4 is an example diplomatic cable (plus metadata) retrieved through a query in our database's CFPF collection.[18]

FIGURE 4: An Example Cable from the Central Foreign Policy File Collection
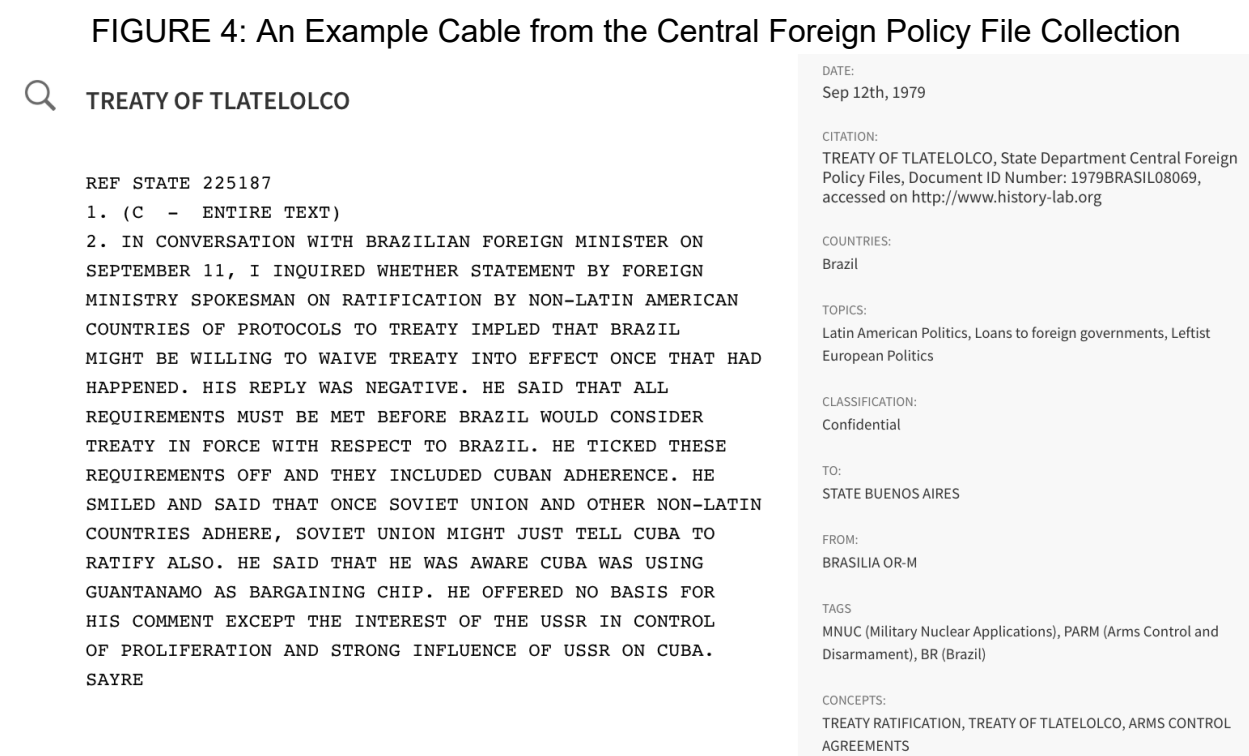
In words, the output above reveals the following: the document's identifier is "1979BRASIL08069" in the CFPF corpus. It is a cable entitled "TREATY OF TLATELOLCO" drafted on September 12th, 1979. It was sent by the US embassy in Brasilia to the State Department headquarters in DC and to the Buenos Aires embassy. It was classified as Confidential and had subject TAGS

---

[17] Additionally, the metadata includes handling instructions such as "limited distribution" and "eyes only".
[18] The document can be viewed at: http://www.history-lab.org/documents/1979BRASIL08069

"MNUC" (Military Nuclear Applications) and "PARM" (Arms Control and Disarmament) as well as a geographic TAG "BR" (Brazil). The cable was also associated with the concepts "TREATY RATIFICATION," "TREATY OF TLATELOLCO," "ARMS CONTROL AGREEMENTS" as relevant to the cable. Not shown in the figure above but included in the database, the cable's designation "ACTION ACDA" tells us that the cable was sent to the Arms Control and Disarmament Agency within the State Department. This cable was sent as a telegram (TE). Its topics (from the topic model) were "Latin American Politics", "Loans to foreign governments", and "Leftist European Politics."

## 4.5. *Foreign Relations of the United States* Collection

The *Foreign Relations of the United States* (FRUS) collection is the next largest corpus in FOIArchive. It is comprised of documents selected by the Office of the Historian of the State Department representing the official record of US foreign relations. The print version consists of more than 450 volumes organized mostly by country or region, but sometimes by a subject such as the Berlin Crisis or Arms Control. Until the volumes on the 1940s, the documents were almost exclusively from the State Department, including presidential meetings and communications recorded by diplomats. But more recent volumes contain a broader selection of documents, including some from the National Security Council, the Pentagon, and the CIA.

The collection spans a long period, with documents ranging from November 3, 1620 to January 2, 1989. Our processed collection currently contains all available volumes. The volumes are organized chronologically into "subseries" by administration, and geographically and topically within each subseries.

The FRUS collection includes various types of documents ranging from private documents previously classified as "Top Secret," "Secret," or "Confidential" to public documents (Figure 5, though note we truncate the time axis simply for display reasons).[19]

---

[19] The FRUS documents do not have the classification in the original metadata, and the ones prior to World War II were not typically classified. Where possible, we extracted the classification for more recent documents from the full text. There are 410 documents from 1620 to 1860 that are not shown in the graph.

FIGURE 5: Number of FRUS Documents by Year
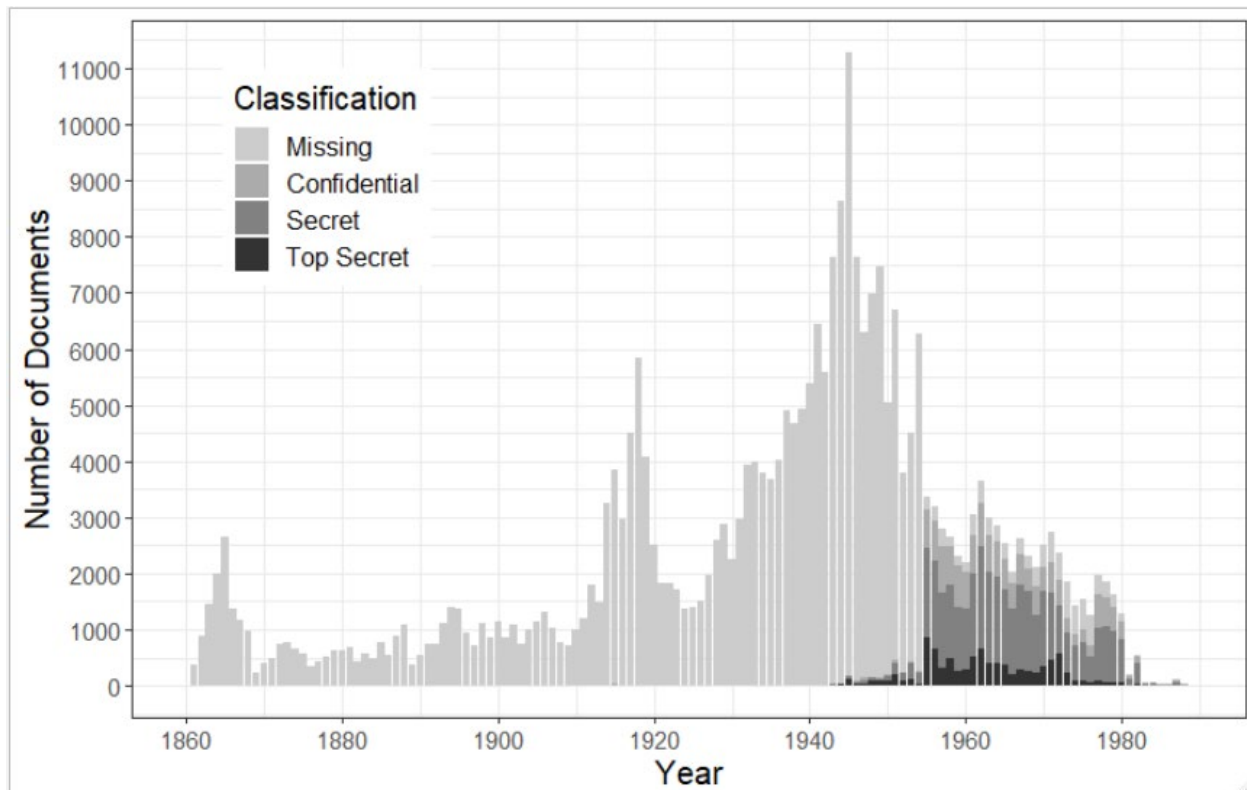
The following variables capture key attributes of the documents' metadata in the FRUS collection: "id," "subject"; "body"; "date"; "classification"; "p_from"; and "p_to."

The variable "id" display a unique identification number for each document. The variables "subject" and "body" are for the document's title and body text, respectively. The variable "date" for this collection shows the relevant date. The variable "classification" represents the sensitivity of the contents of the documents. The variables "p_from" and "p_to" display persons who sent or wrote the document and whom the document is addressing when relevant. About half the FRUS documents include a value for p_from and about a quarter include a value for p_to.

As with the CFPF documents, the following variables about the FRUS documents are also available through our database: "topic," "person," "organization," and "gpe."

In Figure 6 is an (abridged) example document from the FRUS collection retrieved through a query of our database.[20] It was published in the second volume for year 1945 in the FRUS series, with the identifier "frus1945v02d128." The display includes relevant curated topic labels, on the right.

---

[20] The document can be viewed at: http://www.history-lab.org/documents/frus1945v02d128

FIGURE 6: An Example Document from the FRUS Collection
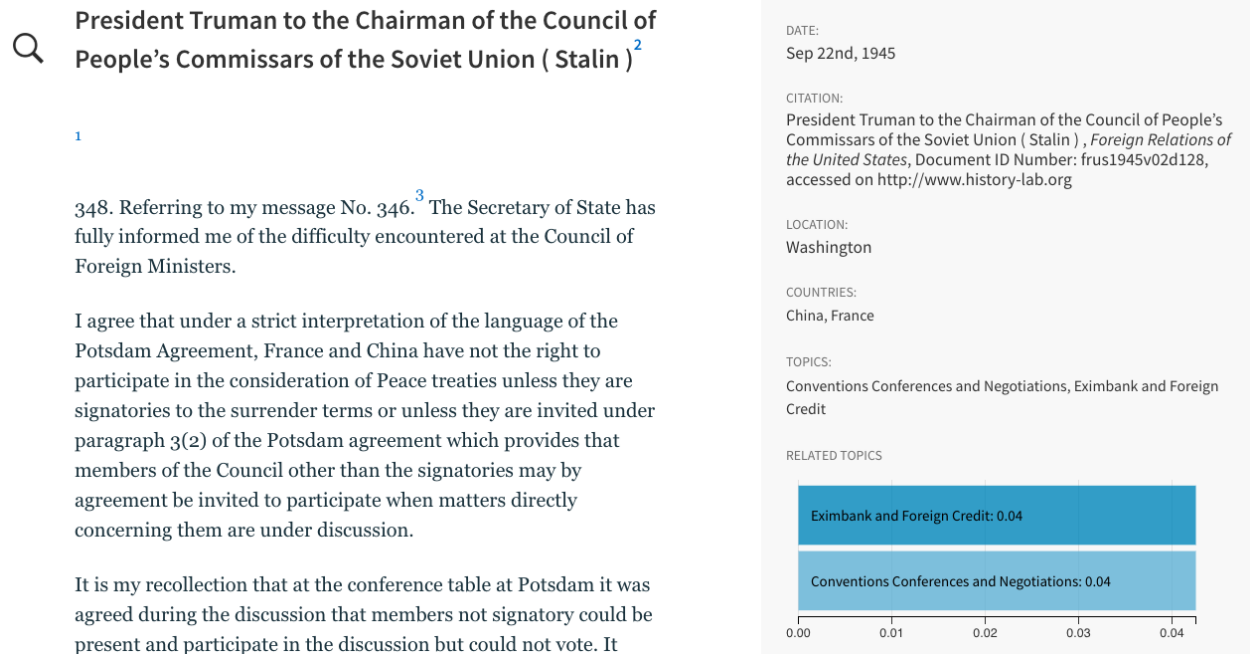
Doubtless, readers can think of many questions our data can be used to answer---including those we noted in the motivation above. But to give a specific example of our collections at work, we now demonstrate how one might address just one question: how to measure "relative importance" in foreign policy.

5. Application: Country TAG Traffic as a Measure of US Diplomatic Priorities, 1973--1979

Now we propose a new measure of US diplomatic priorities based on our CFPF collection. Our purpose is not to produce "the" definitive criterion for this problem (we are, in any case, limited to the period 1973--1979), but rather to show readers the compelling possibilities that such data allow, especially as more and more records in the same collection are released in years to come. As background we define diplomatic importance (or status) as "the relative importance that the states in the system attributed to one another" that is different from power or capability (Small and Singer 1973, 578–79).[21]

The existing literature has pursued several strategies. Small and Singer (1973), for example, count the number of diplomatic missions countries received from the US. Others have studied state voting behavior at the UN General Assembly (see Bailey et al, 2015 for an overview). Though certainly useful, such measures have their own shortcomings. For one, they misestimate

the importance of states that joined the UN later but were nonetheless salient for US policy making.

We complement these efforts, measuring countries' relative diplomatic importance to the US in the 1970s. In particular, we can estimate "fluctuations in a state's diplomatic importance" and "a state's importance in a particular region or in the context of a specific substantive problem," complementing existing efforts to "estimate the more slow-moving importance scores" (Small and Singer 1973, 580).

5.1. Descriptive Statistics of Country TAG Traffic

Our measure is constructed at the country-year level, but can be even more fine-grained depending on a researcher's needs. We count the number of cables tagged for each country per calendar year. Our measure exists for 1,040 country-years and 156 countries from 1973—1979, but it could obviously be expanded if similar data was available for other periods. On average, each country is tagged in 2,545.4 cables in a given year and 17,036.6 cables throughout this period. A country can be tagged in as little 21 cables for a given year and 277 cables in total (Bhutan). It can be tagged in as many as 24,856 cables in a given year and 144,726 in total (the Soviet Union).

Figure 7 gives frequency plots of the measure for non-US contemporary country-years and countries. The distributions are right-skewed, and we see that only a few country-years and countries produce heavy TAG traffic. 85.86% of country-years were tagged in fewer than 5,000 cables. Similarly, 76.92% of countries were tagged in fewer than 25,000 cables.

FIGURE 7: Country TAG Traffic at Country-Year and Country Levels

By Country-Year
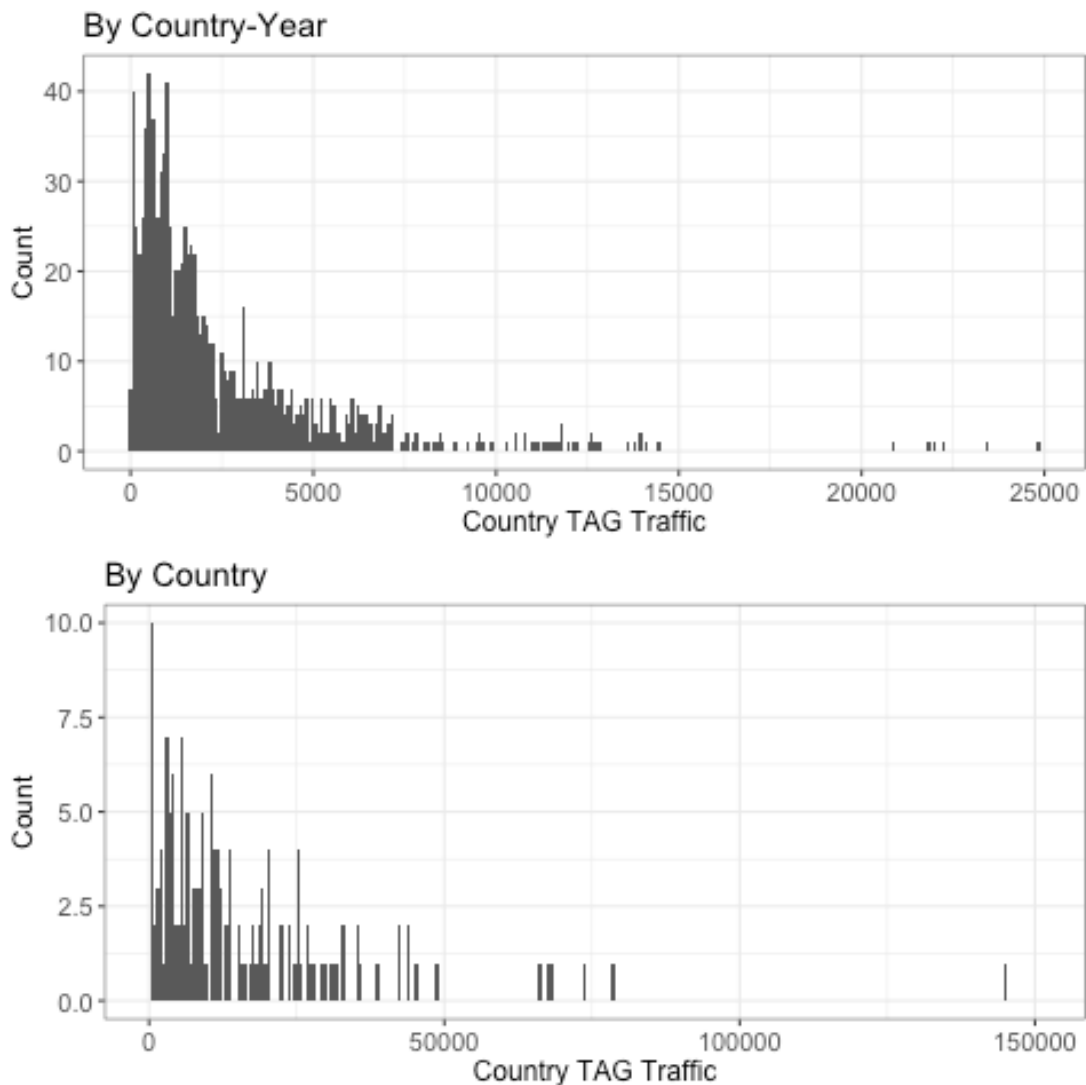
By Country

Note that country TAG traffic is different from cable traffic, and provides a more meaningful measure of U.S. diplomatic priorities. For instance, 20,876 cables included the country TAG for the USSR in 1974. In contrast, only 18,015 cables were sent to or received from the US embassy in Moscow or the Consulate General in Leningrad that year. Moreover, variations in country tag traffic can reflect major events concerning that country even when there is no American embassy to send or receive cables, such as in Tehran during and after the Hostage Crisis. This addresses deficiencies in existing datasets about US diplomacy. For instance, there was no diplomatic representation or presidential visit by the US, and no US-related diplomatic event in the German Democratic Republic (GDR) in 1973-74 or North Vietnam/Vietnam in 1974-79, or Rhodesia in

1978 (e.g. Baggott Carter 2018). However, these countries frequently appeared in our measure for diplomatic importance.

5.2. Validation and Extensions: Using Other Collections

To validate our measure, we look at order statistics. Across all years, the list of 10 countries most frequently tagged in cables reflects US foreign policy priorities in the 1970s. As security threats, the Soviet Union and the GDR were in the first and third positions, respectively. Four developed country allies or economic competitors are in the top 10 including the United Kingdom (2nd) and Japan (4th).

Not surprisingly, these countries also dominate the list of 20 country-years with the largest amount of TAG traffic. In contrast, among the countries least frequently tagged in cables, small countries with few US strategic interests---islands or landlocked countries, in particular---feature prominently.

As an example of a more dynamic story we can tell, consider Figure 8. A natural question for this period is how the retaliatory *Organization of Arab Petroleum Exporting Countries* (OAPEC) embargo of 1973 affected US foreign policy making priorities. In the figure, we report the weekly mentions of OAPEC nations in traffic between US embassies and the State Department from the CFPF collection. We see that the oil shock increased US attention to the region, bringing it to a new, higher plateau than before the crisis (we confirmed this with statistical breakpoint tests). The U.S. relied particularly on Saudi Arabia for information about the crisis, and we can look at other document collections to triangulate our inferences. FRUS document frus1969-76v36d239 from November 1973 suggests that the embargo might be lifted or lessened if the US gave a show of support to the Arab nations. In December 1973, Saudi Arabia reported on the results of an OAPEC conference followed by Kissinger's scathing response a couple of weeks later (frus1969-76v36d265 and frus1969-76v36d273). Finally, a conversation with Bill Donaldson in January 1974 raised the possibility of bilateral negotiations with Saudi Arabia---something we learn from the Kissinger telephone transcripts (0000C606). At a *prima facie* level then, OAPEC was successful in garnering the more focused attention of the United States. Researchers can of course take a deeper dive, and examine the content of the cables, but we leave that for future study.

FIGURE 8: Political and Economic Cables to OAPEC Nations as a Proportion of All Such Cables Sent[22]
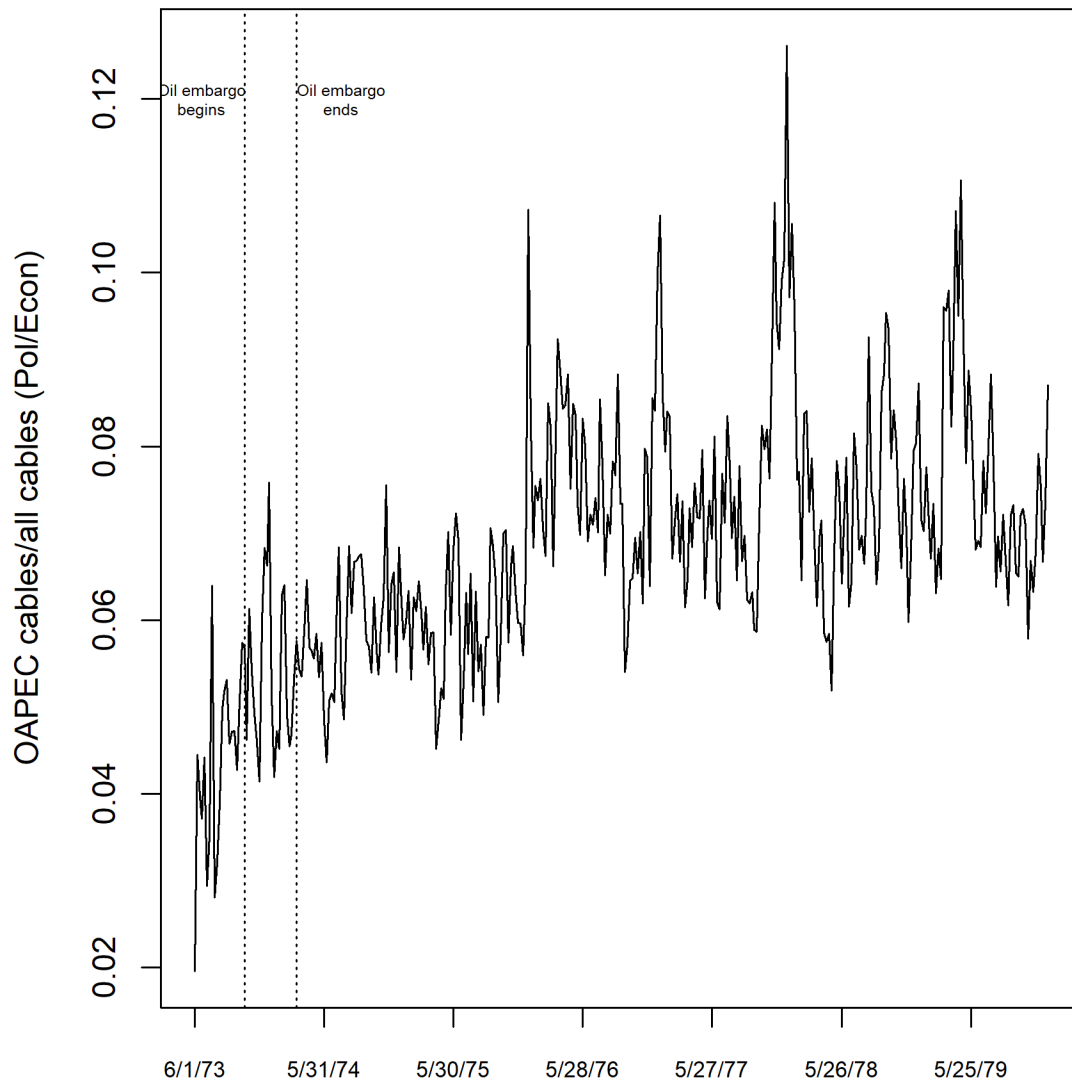
6. Discussion

In this paper, we introduce FOIArchive, a new database of over 3 million documents about diplomacy and foreign policy, especially useful for studying intra-state decision-making. We provide extensive metadata about the documents, including the countries and persons they mention, topics, and classification levels. The metadata includes information we extracted with domain-specific, customized Natural Language Processing tools. The FOIArchive Database also

---

[22] Note the permanent increase in attention after the oil shock of October 1973.

includes an online platform with an API and a website. We showed how it might be used to estimate US diplomatic priorities.

Our data comes with some caveats, as with any new collection. Some texts are unavailable, either because they were not preserved or not released. And, of course, the quality of our metadata is subject to human error. For instance, other than the metadata we extracted ourselves, Central Foreign Policy Files' metadata was created by State Department or US National Archives employees through manual entry, hence it is sometimes incorrect or inconsistent. We strove to detect the errors and verify and clean the metadata.

Future opportunities using the data abound. The internal documents will be a useful resource for scholars interested in private information, signaling, and bureaucratic politics. Our NER data about people, organizations, and countries will be important for researchers to generate and test their arguments about diplomacy at the micro-level. It will provide opportunities for researchers to examine the role of various entities in the US foreign policy making process and to assess their effect on outcomes. More generally, our resource allows millions of records to be (almost) instantly brought to bear on questions of interest to social scientists, historians and others.

References

Baggott Carter, Erin. 2018. *American Diplomacy Dataset*.

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. Research & Politics.

Berridge, G.R. 2015. *Diplomacy: Theory and Practice*. Palgrave Macmillan UK.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (January): 993–1022.

Correlates of War Project. 2017. *State System Membership List, V2016*. http://correlatesofwar.org.

Erkan, Günes, and Dragomir R. Radev. 2004. "LexRank: Graph-Based Lexical Centrality As Salience in Text Summarization." *J. Artif. Int. Res.* 22 (1): 457–479.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88 (3): 577–592.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework*. Cambridge University Press.

Gertz, Geoffrey. 2018. "Commercial Diplomacy and Political Risk." *International Studies Quarterly* 62 (1): 94–107.

Gill, Michael, and Arthur Spirling. 2015. "Estimating the Severity of the WikiLeaks United States Diplomatic Cables Disclosure." *Political Analysis* 23 (2): 299–305.

Gray, Julia, and Philip BK Potter. Forthcoming. "Diplomacy and the Settlement of International Disputes." *Journal of Conflict Resolution*.

Haglund, Evan T. 2015. "Striped Pants versus Fat Cats: Ambassadorial Performance of Career Diplomats and Political Appointees." *Presidential Studies Quarterly* 45 (4): 653–78.

Hall, Todd, and Keren Yarhi-Milo. 2012. "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." *International Studies Quarterly* 56 (3): 560–73.

Holmes, Marcus. 2013. "The Force of Face-to-Face Diplomacy: Mirror Neurons and the Problem of Intentions." *International Organization* 67 (4): 829–61.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. 1st ed. Princeton, NJ: Princeton University Press.

Katagiri, Azusa, and Eric Min. 2019. "The Credibility of Public and Private Signals: A Document-Based Approach." *American Political Science Review* 113 (1): 156–72.

Kinne, Brandon J. 2013. "IGO Membership, Network Convergence, and Credible Signaling in Militarized Disputes." *Journal of Peace Research* 50 (6): 659–676.

Kurizaki, Shuhei. 2007. "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." *American Political Science Review* 101 (3): 543.

Linguistic Data Consortium. 2005. "ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 5.6.1." http://www.ldc.upenn.edu/Projects/ACE/.

Morgenthau, Hans J. 1948. *Politics Among Nations: The Struggle for Power and Peace.* New York: Knopf.

Neumayer, Eric. 2008. "Distance, Power and Ideology: Diplomatic Representation in a World of Nation-States." *Area* 40 (2): 228–36.

Plouffe, Michael, and Roos van der Sterren. 2016. "Trading Representation: Diplomacy's Influence on Preferential Trade Agreements." *The British Journal of Politics and International Relations* 18 (4): 889–911.

Sagan, Scott D. 2014. "Two Renaissances in Nuclear Security Studies." In *H-Diplo/ISSF Forum*. Vol. 2.

Sartori, Anne. 2002. "The Might of the Pen: A Reputational Theory of Communication in International Disputes." *International Organization* 55 (3): 121–149.

Singer, J. David, and Melvin Small. 1966. "The Composition and Status Ordering of the International System: 1815-1940." *World Politics* 18 (2): 236–82.

Small, Melvin, and J. David Singer. 1973. "The Diplomatic Importance of States, 1816–1970: An Extension and Refinement of the Indicator." *World Politics* 25 (4): 577–599.

Trachtenberg, Marc. 2006. *The Craft of International History*. Princeton University Press.

Trager, Robert F. 2017. *Diplomacy: Communication and the Origins of International Order*. Cambridge University Press.

US National Archives and Records Administration. 2019. "Frequently Asked Questions: Record Group 59: General Records of the Department of State Central Foreign Policy File, 1973 – 1979." January 30, 2019. https://www.archives.gov/files/research/foreign-policy/state-dept/rg-59-central-files/faqs.pdf.

Wheeler, Nicholas. 2018. *Trusting Enemies: Interpersonal Relationships in International Conflict.* New York: Oxford University Press.