Diplomatic Documents Data for International Relations: The *Computational and Historical Resources on Nations and Organizations for the Social Sciences* (CHRONOS) Database[1]

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

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

Abstract

We introduce the *Computational and Historical Resources on Nations and Organizations for the Social Sciences* (CHRONOS) Database, a comprehensive collection of over 3 million documents about state diplomacy. Substantively, our database provides opportunities to analyze: previously classified (or publicly unavailable) corpora of internal government documents; raw---often full---text of those documents; and within-country diplomatic records for three specific cases, the US, UK, and Brazil. The full span of the data is 1861-2013, but is mainly from the 20th century. Our database also 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, 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Diplomacy, diplomatic communication, US foreign policy, text-as-data, historical documents

[1] Grants from the National Science Foundation supported this work (award numbers 1637108 and 1637159). We thank Erin Baggott Carter and Jonathan Moyer for generously sharing their data with us.

[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 Researcher, Department of Politics and Center for Data Science, New York University, chs298@nyu.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, mitigated and negotiated by way of diplomacy and diplomats. Unsurprisingly then, scholarship has devoted considerable attention to the subject of these 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. Consequently, scholars face a problem: there is both too much data, and too little. Its sheer size makes it hard to catalog and work with, especially if one wishes to move between aggregate analysis and the inspection of individual cases. But then finding more than one example of a particular phenomenon and applying the scientific method to it can be daunting. Perhaps as a result, work on diplomacy has focused on theory---both qualitative and quantitative---with small numbers of case studies, revisited multiple times. 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 secret government documents. We also introduce an online platform with an application programming interface (API) and website with which scholars can search for, download, and analyze datasets customized to their research needs, including the full text of the documents.

This database, *Computational and Historical Resources on Nations and Organizations for the Social Sciences* (CHRONOS), aims to support the text-as-data approach to research on International Relations (IR). It allows scholars to study multiple aspects of diplomacy, especially US diplomacy. This includes adversarial and non-adversarial relations, diplomatic actions and the information conveyed by them in both public and private spheres, and diplomacy by multiple actors, including leaders, rank-and-file diplomats, and different government agencies. 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.

This article proceeds in four sections. First, we discuss prior research on diplomacy and our motivation for creating the database and platform. Second, we describe our contribution to the broader IR literature. Third, we describe our data, with particular focus on two collections. Fourth, to demonstrate the usefulness of our collections, we introduce and validate a new measure of US diplomatic priorities.

2. Motivation: Importance and Complexity of Diplomacy in International Relations

The study of diplomacy in International Relations has given rise to a complex, multifaceted literature that may be nonetheless organized around three broad themes: adversarial and non-adversarial diplomacy, diplomacy as a manipulatable "signal" and a non-manipulatable "index," and diplomacy at the macro- and micro-levels. At a high level, our data collection gives researchers tools and resources to speak to all of these areas. To understand this motivation in greater detail, we now review the major themes in this literature.

Diplomacy can take place among "foes" and "friends" alike. Accordingly, scholars have discussed adversarial and non-adversarial diplomacy. For instance, Morgenthau (1948) viewed diplomacy as "practical statecraft" comprised of the threat of force, persuasion, and compromise. He argued diplomacy must involve correct assessment of a state's own objectives and others' but also emphasized that states must be willing to compromise on all unimportant issues. In contrast, the voluminous literature on coercive bargaining (e.g. Bueno de Mesquita, Morrow, and Zorick 1997; Fearon 1994, 1995, 1997, 2013; Leventoglu and Tarar 2008; Powell 2004; Slantchev 2003a, 2003b), including works focusing on diplomacy per se (Guisinger and Smith 2002; Kurizaki 2007; Lindsey 2017; Sartori 2005; Trager 2017), has highlighted adversarial diplomacy (or its futility) in inter-state disputes over security issues (Trager 2016) and economic ones (e.g. Bayne and Woolcock 2016; Gray and Potter 2017).[7] Many authors have discussed nation-states' direct communication with other enemy states or indirect communication with the adversaries---such as communication with allies or mediators in the presence of adversaries in crisis bargaining "because communication is thought to be most difficult in such contexts" (Trager 2016, 220).

Of course, diplomacy need not be adversarial (e.g. Bull 1977; Watson 1984) and in that context, scholars have analyzed the interaction between domestic politics and diplomacy (Putnam 1988), forum shopping (Pekkanen, Solís, and Katada 2007), or mediation (Fey and Ramsay 2010; Kydd 2003), for instance. This is complemented by a literature on diplomatic visits that:

[7] These literatures encompass both inter-state disputes over security issues and those over economic issues. However, "the influence of diplomacy in economic disputes remains understudied, particularly quantitatively," notes Gray and Potter (2017, 5). While some argue that economic diplomacy is (conditionally) effective in resolving inter-state economic disputes (Bayne and Woolcock 2016; Gray and Potter 2017), others do not. For instance, Tomz (2012) argues in his quantitative and detailed historical analyses about the unimportance of "gunboat" diplomacy in payment of sovereign debts. Scholars have also examined the relationship between adversarial diplomacy and international institutions in trade (e.g Carnegie 2014, 2015).

examines the drivers of US leadership visits (Lebovic 2018; Lebovic and Saunders 2016) and the effect of US diplomatic visits on public opinion abroad (Goldsmith and Horiuchi 2009); uses the Chinese leadership's travel pattern (or those of the Dalai Lama) to measure Chinese diplomatic priorities (Fuchs and Klann 2013; Kastner and Saunders 2011; Li 2015); and analyzes US and non-US diplomatic visits to protégé states (McManus 2018).

Beyond diplomatic "actions" themselves, scholars have thought theoretically about the information---and the effects of that information---communicated and conveyed by diplomacy.[8] Within this context, researchers have analyzed diplomacy as a "signal" that is strategic and manipulatable by states. This is as opposed to a reliable "index"---an unintentional act that is impervious to deception or manipulation of the participating states (Jervis 1970, 2017a). In this world, communication involves not only public, but also private information. Here, investigations have centered on diplomacy's validity and effectiveness---broadly, whether diplomatic communication made in private matters for international politics---and the credibility of the information it conveys, that is, when, how, and why diplomacy can be credible in coercive bargaining (Acharya and Ramsay 2013; Carson 2016; Kurizaki 2007; Lindsey 2017; Ramsay 2011; Sartori 2005; Trager 2017).[9]

For instance, the sizable scholarship on crisis bargaining posits that diplomatic signals can be credible when conveyed in public because backing down (or retracting them) is domestically or internationally costly for the sender. The domestic audience in a democracy can punish a leader who backs down by not re-electing him for "violating the national honor" (Fearon 1994), for being incompetent (Slantchev 2006; Smith 1998), or for losing the bargaining benefits from prenegotiation public commitments (Leventoglu and Tarar 2005). Alternatively, the international audience can punish a state or a leader for bluffing by not granting credibility to future diplomacy of the country (Sartori 2002, 2005) or its leader (Guisinger and Smith 2002). The empirical literature has also focused on diplomatic communication and its role as a signal, measuring resolve from diplomatic statements (Katagiri and Min 2019; McManus 2017a) or studying previously private documents of the US or UK governments (Gill and Spirling 2015; Katagiri and Min 2019; Renshon 2009; Trager 2017), for instance.

The backdrop to this literature is questions about states' beliefs and their effect on international political outcomes. Starting at least with Schelling (e.g. 1966, 1980), scholars have argued that despite (or perhaps because of) the anarchical structure of international politics and the ensuing security dilemma, states' beliefs and perceptions are a variable, not a constant, that can affect

[8] Similarly, the literature on covert military action (Carson 2016, 2018; Carson and Yarhi-Milo 2017) explores its informational implications.

[9] The focus on validity and credibility of diplomacy in coercive bargaining largely stems from many rationalist scholars' implicit or explicit comparison of private, diplomatic communication with public communication set as the baseline (Katagiri and Min 2019). Implicitly, these scholars were arguing for or against the view that diplomacy consists of only non-costly signals among actors with diverging interests, hence meaningless and ineffective "cheap talk" (Crawford and Sobel 1982; Farrell and Gibbons 1989).

cooperation (Jervis 1976, 2017b). Consequently, states, or their leaders per se, are often susceptible to being misled or deceived by others' (Jervis 1970, 2001, 2017a). Ultimately then, they may possess inaccurate information and perceptions about the status quo and their adversary's intention. Thus they fail to predict key political events, resulting in fatal intelligence failures (Jervis 2011). Whether and how diplomacy affects these beliefs remains important in the political science literature.

The above notwithstanding, treating diplomacy as an observable "index" (Jervis 1970, 2017a) has also generated much work. For instance, scholars have investigated diplomatic representation (Bayer 2006; Moyer, Bohl, and Turner 2016a; Singer and Small 1966; Small 1977; Small and Singer 1973), explored the formation of diplomatic networks (Kinne 2013b; Maliniak and Plouffe 2011; Neumayer 2008; Xierali and Liu 2006), and analyzed the effect of bilateral diplomatic ties on formation of preferential trade agreements (Plouffe and van der Sterren 2016). This is complemented by the aforementioned literature on diplomatic visits that analyzed US leadership visits (Goldsmith and Horiuchi 2009; Lebovic 2018; Lebovic and Saunders 2016), the Chinese leadership's travel pattern (or those of the Dalai Lama) (Fuchs and Klann 2013; Kastner and Saunders 2011; Li 2015), or both US and non-US diplomatic visits (McManus 2018). Researchers have explored outcomes of diplomatic activities on international policy outcomes, such as the effect of the US government's diplomatic exchanges of gifts on its settlement of international trade disputes at the World Trade Organization (Gray and Potter 2017) or the effect of diplomatic interventions on civil war outcomes (Regan, Frank, and Aydin 2009). With respect specifically to "text" data, scholars have studied public statements by US Secretaries of State (Cogburn and Wozniak 2013), International Criminal Court-related statements (Boehme 2018), speeches about global economic and human rights institutions made at the United Nations General Assembly (Kentikelenis and Voeten 2018), and US presidential statements (McManus 2014, 2017a, 2017b) including public speeches made by President George W. Bush (Renshon 2008).

As with IR literatures on other topics, scholarship on diplomacy has shifted from analysis at the macro-level, such as diplomacy among great powers (3rd image), to the micro-level (1st and 2nd images) including the diplomatic behavior of policymakers and the domestic politics of diplomacy. In addition to studying prominent historical cases, many scholars have built their theories about diplomacy from the "ground-up", sometimes testing them on individuals---including elites---in experimental settings. This includes the literature on face-to-face diplomatic communication among leaders (Hall 2015; Hall and Yarhi-Milo 2012; H olmes 2013, 2018; Rathbun 2014; Yarhi-Milo 2014; Wheeler 2019). Scholars have also studied the connection between diplomatic (usually public) events or visits and US leaders' domestic political incentives and constraints (Baggott Carter 2018b; Potter 2013). Finally, the causes and consequences of diplomatic appointments have been of interest. Here, researchers have focused on the distinction between political and career appointments for US ambassadorships (Arias and Smith 2018; Haglund 2015; Hollibaugh 2015) and examined the effect of ambassadorial vacancies on commercial diplomacy (Gertz 2018).

As this review suggests, the growing body of research on diplomacy has greatly advanced the study of international relations. Nonetheless, several methodological problems remain. For one, IR scholars find it difficult to obtain comprehensive data that would best test their theories. 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 begun to declassify millions of electronic records, but without any easy way to download the underlying data for quantitative analysis, or the finding aids and expert archivists that have long supported qualitative research. 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 CHRONOS Database Provides

The CHRONOS Database provides a free and open platform giving researchers the ability to analyze diplomacy by data-mining a large body of heretofore private information (i.e., classified information). It will support different approaches to IR research, but could also help bring them together. Because the data is extracted from documents-–and linked to those documents–-researchers can shift from an aggregate view of the work product of large organizations to a micro-view of individual documents that produced particular data points. Both quantitative and qualitative research can therefore become more transparent, rigorous, and replicable.

Our datasets are focused on---but not exclusively about---US foreign policy and diplomacy. They include data on both adversarial and non-adversarial diplomacy, private and public information and actions, diplomacy at multiple levels of analysis, and communication across diverse topics. For example, our data includes metadata and/or the full text of the Central Foreign Policy Files, a corpus of over 3.2 million diplomatic documents from the US government. The communications include crisis bargaining, but also the baseline data representing more typical everyday activity. We offer a rare scholarly opportunity to integrate and compare both on the same spectrum, and new data will be added with each yearly release of newly-declassified documents.

Additionally, our data is novel in that it displays almost the entire informational flow among the various levels of the US government’s foreign policy apparatus. In particular, our data sheds light on private information, crucial in the literature on crisis bargaining, by providing previously classified internal government documents. More broadly, our data can also enlighten scholars about diplomatic actors’ preferences and choice of actions regarding private as well as public information and signalling. Our trove of classified documents can also shed light on states’ covert action from a diplomatic perspective.

Our corpora allow researchers to study diplomacy at multiple micro-levels of analysis by providing high-quality information about actions, beliefs, knowledge, and preferences of multiple diplomatic actors, such as Presidents, National Security Advisors, Secretaries of State, and rank-and-file diplomats, as well as actions and decisions undertaken by agencies. We do this via 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 standalone specialized systems that have high accuracy for certain types of high value entities, such as countries, that conventional implementations miss. We also identify and extract Geographical/Social/Political Entities (GPE) which are “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). In sum, our system is specialized in the domain appropriate for IR research, having been trained with government documents about foreign policy.

These resources will contribute to the recent investigations of the micro-foundations of international politics. Specifically, our granular data can extend or test existing arguments about individuals in foreign policymaking. For instance, existing scholarship about leaders and their perceptions, including threat perceptions, strategic rationality, resolve, status concerns and intentions, can be extended to diplomats or foreign policy elites. Our data also allows the investigation of the broader role of foreign policy elites in international politics in a way that complements earlier research that mostly uses survey experiments or case studies (Guisinger and Saunders 2017; Saunders 2015, 2017).

Our data can generate new scholarship about the role of foreign policy elites or bureaucratic agencies in foreign policymaking. For good reason, existing empirical work tends to focus on the effect of individual---usually policymaker/leader---level attributes that function as country-level attributes on foreign policy outcomes of one country, usually the US (e.g. Arias and Smith 2018; Gertz 2018; Gray and Potter 2017) or consist of analyses at the cross country level (e.g. Kinne 2013; Maliniak and Plouffe 2011; Neumayer 2008; Plouffe and van der Sterren 2016; Regan, Frank, and Aydin 2009; Xierali and Liu 2006). While it is unclear what effect, if any, bureaucratic agencies or structure in foreign policy have on foreign policy outcomes or events, it is worth exploring further (Drezner 2000). For instance, Yarhi-Milo (2013) argues that intelligence organizations missions and practices “strongly shape which information they will regard as informative” for making inferences about adversaries’ intentions. (Yarhi-Milo 2014, 261). Similarly, noting fragmentation of foreign aid, scholars have suggested that its allocation or aid delivery are affected by the government’s bureaucratic design (Arel-Bundock, Atkinson, and Potter 2015) or structure (Dietrich 2016; Kilby 2011). The metadata for the Central Foreign Policy Files identifies the offices included---or excluded---for particular communications, allowing researchers to explore inter-office dynamics both within the State Department and vis-a-vis other departments and agencies. And because CHRONOS includes multiple corpora, researchers will also be able to compare and contrast the work product of different bureaucratic structures, such as the CIA and the Pentagon.

Furthermore, IR scholars can use our NER data to study the role of reputation and address the chasm between the views of scholars and those of practitioners (cf. Renshon, Dafoe, and Huth

2018). Many scholars view reputation as attaching to states (Yarhi-Milo and Weisiger 2015) or leaders (Yarhi-Milo 2018) whereas practitioners view reputation as attaching to individual diplomats (Trager 2016, 212).[10] Similarly, practitioners place much more emphasis on the role of “chemistry” in diplomatic relations—which is largely ignored by academics.[11] Our datasets’ NER variables distinguish between individuals, countries, and GPEs mentioned in the documents, allowing scholars to empirically test various hypotheses about the reputation of states, leaders, and diplomats.

More generally, we hope our NER efforts will inspire follow-up computational analyses of historical documents in IR research. Automated parsing of and extraction of information from historical documents with computational tools have been implemented (Boehme 2018; Cogburn and Wozniak 2013; D’Orazio et al. 2014; Katagiri and Min 2019; Kentikelenis and Voeten 2018; Lindsey and Hobbs 2015; Palmer et al. 2015) but generally under-utilized by scholars of IR and US foreign policy.[12]

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 LDA model relies upon word co-occurrence patterns within documents to uncover specialized probability distributions–referred to as “topics”–over the set of unique words represented in the corpus. Each topic assigns high probabilities to words that tend to occur together within that theme. In contrast to traditional document clustering methods, in which a document must be associated with only one topic, the LDA model allows a document to be associated with multiple topics in varying proportions.

While the techniques can be powerful, labeling output from such models (that is, 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 have addressed these shortcomings by having historians curate our topic modeling results. They examine both high-ranking and randomly selected documents for each topic to determine whether they collectively represent a historically-meaningful topic. If so, they assign a label to the topic. For example, documents with a high proportion of the words “bank,” “fund,” “review,” and “rate” belong to the topic labeled “International Development Loans.” Non-meaningful topics are not assigned a topic name but the topics are retained in the database.

4. The CHRONOS Database

In this section, we describe the documents that constitute our data stored in the CHRONOS Database. We then discuss the process of collecting, processing, and integrating the

[10] See Dafoe, Renshon, and Huth (2014) for a review of the international security literature on reputation and status.

[11] For example, negotiations between the US and Iran might have gone better because Kerry was not phased by Zarif’s behavior (personal communications).

[12] See Allen and Connelly (2016) for a discussion of adaptation of computational tools in history.

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

Our data consists of 8 groups of government documents (corpora or “collections”): six corpora of US government documents and one collection each from the UK and Brazil. US corpora include the documents from the State Department *Central Foreign Policy Files*, the *Foreign Relations of the United States* series, the *President’s Daily Briefs*, the *Henry Kissinger Telephone Conversation Collection*, the *Hillary Clinton Email Collection*, and the *Declassified Documents Online* collection. Our database also includes the *Cabinet Papers* from the UK and the *Azeredo da Silveira Papers* from Brazil.

Our *Central Foreign Policy File* (CFPF) 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.[13] This corpus was parsed from the source XML (Extensible Markup Language) files provided by the US National Archives.[14]

Our data also includes 209,046 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 1861--1985. This corpus was also parsed from XML files downloaded from the State Department’s GitHub site.[15]

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.

[13] 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).

[14] The CFPF collection will be discussed in more detail later in the paper.

[15] Details of the FRUS collection will also be discussed later in the paper.

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.

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 and FBI in addition to the Department of State. We parsed the source XML files and images provided by Gale Cengage to create our collection.

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.[16] 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 CHRONOS. It also includes documents in multiple languages, a test-bed for research in which language, e.g. translations, is of interest.

Altogether, CHRONOS currently includes 3,657,378 documents. Of these, we provide full text of over 2.9 million documents.[17] Our data also covers a long time span: the oldest document dates

[16] This paper focuses on US government documents, however.

[17] 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).

from 1861, the newest is from 2013. But the collection is richest for the 1970s, the beginning of the era of electronic records, and it will continue to grow as more records are reviewed and released in years to come. Below 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)	May 2, 1861-- Apr. 5, 1985	209,046	209,046	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			3,604,560	2,927,172	677,388	
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			3,657,378	2,979,990	677,388	

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, which provided us with the documents, if users would like to see the original document.

4.2. Workflow

CHRONOS is composed of documents that we obtained from a source, processed, and ingested. For each collection, we set up three work-flows: pre-processing, processing, and post-processing. In the pre-processing stage, we scraped the source documents from their websites, saved them 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. This stage also involved determining which variables to extract and how, and creating and testing different parsers (scripts) for extracting the variables customized to each collection and variable. We cleaned the extracted text, revising its formatting or removing unnecessary metadata, for instance. In this stage, the input was the source documents in XML or PDF format and the output was digitized text.

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 used the SQLAlchemy library in Python to store the pre-processed data in our MySQL database. 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. The website also offers graphical representation of the data based on the SQL tables. 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. Historians examined both high-ranking and randomly selected documents for each topic to determine whether they collectively represent a historically-meaningful topic. If so, they assigned a label to the topic. We also reconciled entity names and other metadata to correct for typos or other errors.[18]

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 as new documents are added.

[18] To ensure compatibility with other sources we keep the metadata with errors as well.

4.3. Online Platform: API and Website

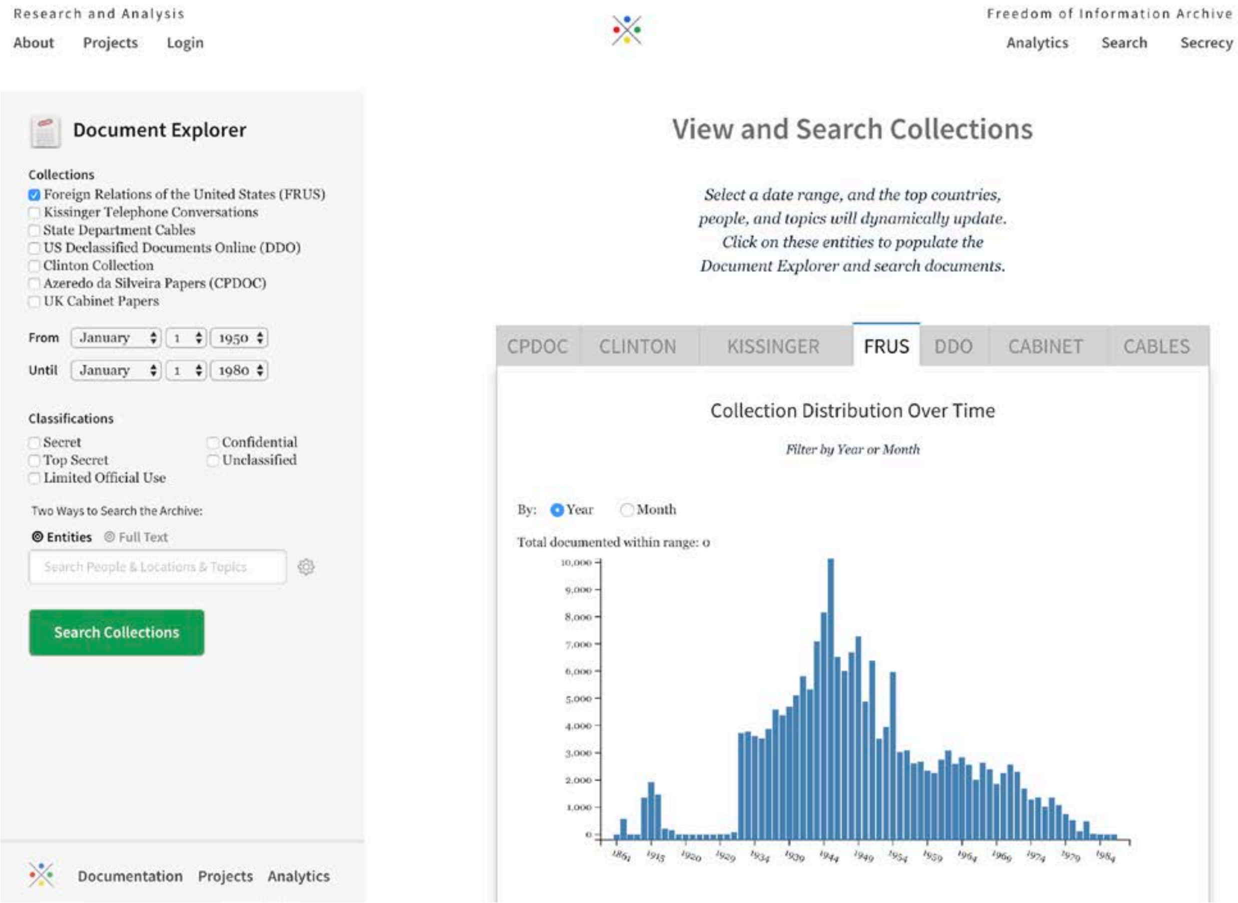
Our database makes all of the documents and associated metadata accessible through an Application Programming Interface (API) and on the website. By using unique, encrypted identification keys obtained by free registration, researchers can use the API to download customized data in JSON format, including full-text of the documents.[19] We have also created Stata and R interfaces with the API to allow users to download directly into either statistical package.[20]

The CHRONOS data is also viewable via a website (http://www.history-lab.org/). This version, named the “Freedom of Information Archive,” predates CHRONOS, and is designed and maintained to make the same data accessible to a broader public through a graphical interface. Researchers can explore all the metadata and the documents using filtering parameters and interactive tools. Below is a screenshot of the Document Explorer of the website to which users can configure and submit their queries and view the results:

[19] The API is built on industry-standard Representational State Transfer (RESTful) practices. It is set up for Cross-Origin Resource Sharing and returns information in a format consistent with the US government’s Open Data Initiative.

[20] Links to the Stata and R packages will be available on our website (history-lab.org). The website will also have a document with more detailed instructions for using the API.

FIGURE 1: Screenshot of Our Website



Document Explorer

Collections

☑ Foreign Relations of the United States (FRUS)
□ Kissinger Telephone Conversations
□ State Department Cables
□ US Declassified Documents Online (DDO)
□ Clinton Collection
□ Azeredo da Silveira Papers (CPDOC)
□ UK Cabinet Papers

From January 1 1950
Until January 1 1980

Classifications

□ Secret
□ Top Secret
□ Limited Official Use
□ Confidential
□ Unclassified

Two Ways to Search the Archive:

● Entities ○ Full Text

Search People & Locations & Topics

Search Collections

View and Search Collections

Select a date range, and the top countries, people, and topics will dynamically update. Click on these entities to populate the Document Explorer and search documents.

Collection Distribution Over Time

Filter by Year or Month

By: ● Year ○ Month

Total documented within range: 0

The interface includes tools to view the relative distribution of documents over time in each collection. Users can also view the most important topics, and the most frequently mentioned countries and persons. By adjusting the date range, they can view the topics, countries, and persons most prominently featured in particular periods. They can also search the database for documents sharing these features. The Merriam tool will automatically retrieve similar documents from all collections based on specified criteria, such as simultaneity and common semantic features. A separate analytics page permits users to further explore the topic-modeling data, including the quantitative data on the words that make up a topic, the most relevant document for each topic, and topic relevance scores for each document.

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.

These broad tools described, we now turn to describing our two largest collections in some detail.

4.4. Central Foreign Policy File Collection

The *Central Foreign Policy Files* (CFPF) collection is the biggest corpus in CHRONOS. The collection is comprised of declassified 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) from 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, including all those concerning political matters. They have not preserved records related to passports and visas or administrative matters 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 2013a, 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).[21]

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. Descriptive statistics of these variables are included in Appendix A.

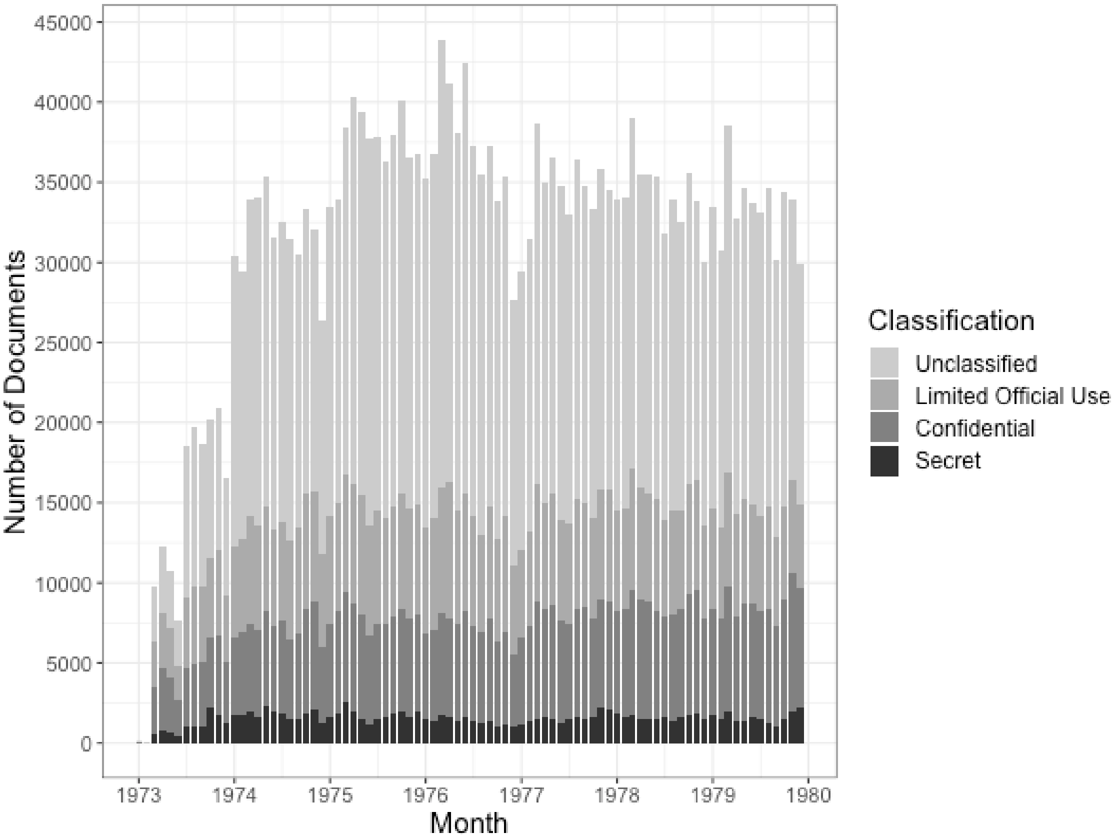
The following variables constitute key metadata of the documents in this corpus: id; subject; body; date; classification; from; to; tag; concept; office; and type. The variable “id” represents a unique identification number for each cable. The variables “subject” and “body” are for the

[21] In contrast, US embassy reporting on Iran’s domestic politics and the Shah prior to the Iranian Revolution left the US government largely unprepared for the Revolution “partly because of its lack of contacts with non-elite groups” (Jervis 2011, 64).

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.[22] 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.

Below we display the number of cables in the CFPF corpus by month and classification level (FIGURE 2). The corpus includes over 3 million cables previously classified “Secret,” “Confidential”, “Limited Official Use” or “Unclassified.”

FIGURE 2: 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

[22] 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.

predefined fields (values) by the State Department that corresponds to a variety of subjects. Each cable will have some number of TAGS. The major categories include:

- Geographic TAGS: All countries and territories at the time are represented by a country TAG (e.g. AF for Afghanistan, IN for India).
- Organization TAGS: This category includes a variety of organizations ranging from the very important to the more mundane, such as "ADF" for the African Development Fund, "AEC" for the Atomic Energy Commission, "CPSU" for the Communist Party of the Soviet Union, "DOD" for the Department of Defense, and "AER LINGUS" for Irish Airlines.
- Programs TAGS: These include such TAGS as "KNNP" for "Nuclear Non-Proliferation," "KTIA" for "Treaties and International Agreements," "KSTC" for "Strategic Trade and Technology Controls," and "KSUM" for summit meetings.
- Subject TAGS: These are broad "subject" areas that cables address, such as "Administration," "Business Services," "Consular Affairs," "Economic Affairs," "Military Affairs," "Outreach," "Political Affairs," "Social Affairs," and "Technology/Science".[23]

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. For example, the most popular concepts are "MEETINGS," "VISITS," "NEGOTIATIONS," "POLICIES," "REFUGEES," "AGREEMENTS," "FOREIGN RELATIONS," "DIPLOMATIC DISCUSSIONS," and "FOREIGN POLICY POSITION."[24]

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.

[23] Within each of these subjects are additional subfields, such as "ABUD" for "Administration – Budget Services and Financial Systems," "BENC" for "Business Services – Engineering and Construction Services," "CPASS" for "Consular Affairs – Passports and Citizenship," "ECRE" for "Economic Affairs – Construction, Repair, and Expansion," "MASS" for "Military Affairs – Military Assistance and Sales," "OPDC" for "Outreach – Diplomatic Correspondence," "PBTS" for "Political Affairs – Boundary and Sovereignty Claims," "SHUM" for "Social Affairs – Human Rights," and "TPHY" for "Technology/Science – Physical Sciences".

[24] However, there are some concepts that are more specific but do not appear nearly as frequently, such "SURPLUS WEAPONS DISPOSAL," "RELIGIOUS DISCRIMINATION," "COMMUNIST DOCTRINE," and "ELECTION INTERFERENCE." There are also concepts which refer to the mundane, everyday work the State Department engages in which include concepts like "PRICES," "DIPLOMATIC PERSONNEL," "PERSONNEL TRAINING," "PERSONNEL TRAVEL," and "HOTEL ACCOMODATIONS."

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,” “GC” for a “General Correspondence,” “IR” for an “Intelligence Report,” “ME” for a “Memorandum,” “MC” for a “Memorandum of Conversation,” “OM” for an “Operations Memorandum,” and “TE” for a “Telegram”.

In addition, we have also created the following variables by processing the body text of the cables: “topic,” “person,” “organization,” and “gpe.” The variable “topic” refers to the topics discussed in 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. We classify each mention of GPEs into four types---GPE as an organization, GPE as a person, GPE as a location, and GPE as a GPE.[25]

Below is an (abridged) example diplomatic cable retrieved through a query in our database with the variables noted in parentheses.[26] We also filled in the description of each TAGS field.

[collection] statedeptcables
[id] 1976ECBRU06967
[subject] COMMISSION PROPOSAL TO TAX ELIGIBLE VEGETABLE/ MARINE OILS
[body] 1. SUMMARY: FOLLOWING PRESENTATION ON JULY 12 OF SUBSTANCE
OF STATE 170962 TO HANS WIJNMAALEN, CHIEF OF CABINET OF
EC COMMISSIONER FOR AGRICULTURE LARDINOIS, WIJNMAALEN GAVE
AGATTACHE MONTEL AND FRASER (AGRIC) NEW INFORMATION ON
COMMISSION'S PRESENT THINKING REGARDING APPLICATION OF PROPOSED
TAXES ON MILK AND EDIBLE VEGETABLE/MARINE OILS.
END SUMMARY.
2. WIJNMAALEN, IN RESPONSE TO QUESTIONS BY MONTEL AND
FRASER, EXPLAINED THAT PROCEEDS OF PROPOSED TAX ON MILK WOULD
BE PAID INTO FEOGA FUNDS. THIS IS AUTHORIZED UNDER ARTICLE
43 OF ROME TREATY.
…
HINTON
[date] July 14th, 1976

[25] For instance, France is tagged as a GPE as an organization in the sentence “France signed a treaty with Germany last week.” France is a GPE as an organization in the sentence “France vacations in August” and a GPE as location in the sentence “The world leaders met in France yesterday” (Linguistic Data Consortium 2005, 15).

[26] The document can be viewed at: http://www.history-lab.org/documents/1976ECBRU06967

```
[classification] Limited Official Use
[to] STATE
[from] EC BRUSSELS
[tag] EAGR (Agriculture and Forestry), EPAP (Plant, Animal, and Wood Products), EEC (European Common Market (European Communities))
[concepts] POLICIES, CUSTOMS DUTIES, VEGETABLE OILS, AGRICULTURAL COMMODITIES
[office] ACTION EUR
[type] TE
[topic] Agricultural productivity estimates, International financial developments, Natural resource distribution data
```

In words, the output above reveals the following: the document's identifier is "1976ECBRU06967" in the CFPF corpus. It is a cable entitled "COMMISSION PROPOSAL TO TAX ELIGIBLE VEGETABLE/ MARINE OILS" drafted on July 14th, 1976. It was sent by the US mission to the European Commission in Brussels to the State Department headquarters in DC. It was classified as Limited Official Use and had subject TAGS "EAGR" (Agriculture and Forestry) and "EPAP"(Plant, Animal, and Wood Products) as well as an organization TAG "EEC" (European Common Market (European Communities)). The cable was also associated with the concepts "POLICIES," "CUSTOMS DUTIES," "VEGETABLE OILS," and "AGRICULTURAL COMMODITIES" as relevant to the cable. The cable's designation "ACTION EUR" tells us that the cable was sent to the Bureau of European and Eurasian Affairs within the State Department. This cable was sent as a telegram (TE). Its topics (from the topic model) were "Agricultural productivity estimates", "International financial developments", and "Natural resource distribution data."

4.5. *Foreign Relations of the United States* Collection

The *Foreign Relations of the United States* (FRUS) collection is the next largest corpus in CHRONOS. 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.

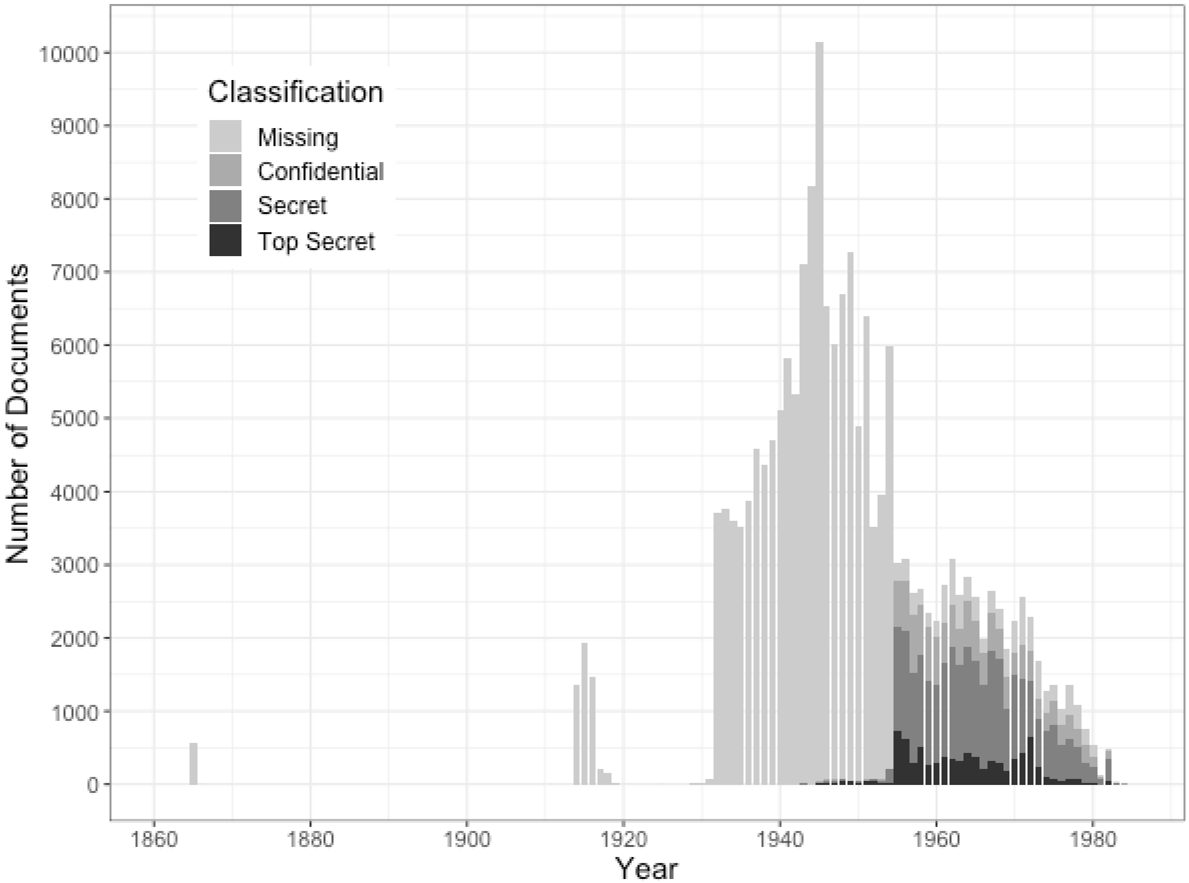
The collection spans a long period, with documents ranging from May 2, 1861 to April 5, 1985. Our processed collection currently contains volumes from 1930--1980, including volumes from 1861 and 1980--1984. The volumes are organized chronologically into "subseries" by administration, and geographically and topically within each subseries. For example, the subseries for the Richard M. Nixon and Gerald R. Ford Administrations (1969--1976) includes the following volumes:

- Volume I entitled "Foundations of Foreign Policy, 1969–1972"

- Volume III entitled “Foreign Economic Policy; International Monetary Policy, 1969–1972”
- Volume VI entitled “Vietnam, January 1969–July 1970”
- Volume XVI entitled “Soviet Union, August 1974–December 1976”
- Volume XVII entitled “China, 1969–1972”
- Volume XXVI entitled “Arab-Israeli Dispute, 1974–1976”
- Volume XXXII entitled “SALT I, 1969–1972”
- Volume XXXV entitled “National Security Policy, 1973–1976”
- Volume XXXVI entitled “Energy Crisis, 1969–1974”

The FRUS collection includes various types of documents ranging from private documents previously classified as “Top Secret,” “Secret,” or “Confidential” to public documents (FIGURE 3).[27]

FIGURE 3: 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.”

[27] 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.

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.[28]

As with the CFPF documents, the following variables about the FRUS documents are also available through our database: “topic,” “person,” “organization,” and “gpe.” The variable “topic” refers to the topics discussed in the body text of each FRUS document that were then validated and curated by historians. The variables “person,” “organization,” and “gpe” refer to the Named Entities we have recognized, extracted, and classified from the text of the FRUS documents.

Below is an (abridged) example document from the FRUS collection retrieved through a query of our database.[29] It was published in the second volume for year 1945 in the FRUS series, with the identifier “frus1945v02d128.”

[collection] frus
[volume_id] frus1945v02
[id] frus1945v02d128
[title] President Truman to the Chairman of the Council of People’s Commissars of the Soviet Union (Stalin)[2]
[body][1]
348. Referring to my message No. 346.3 The Secretary of State has fully informed me of the difficulty encountered at the Council of Foreign Ministers.

…
Can't we agree to regard the unanimous action of the Council on the opening day as an invitation to France and China to participate under the Potsdam Agreement? This is too small a matter to disrupt the work of the Council and delay progress towards peace and better understanding.

Truman

[date] Sep 22nd, 1945
[classification]
[location] Washington
[p_from] Truman
[p_to] Stalin
[country] China, France
[topic] Conventions Conferences and Negotiations; Eximbank and Foreign Credit

[28] See Appendix B for additional details about the FRUS corpus’s metadata.
[29] The document can be viewed at: http://www.history-lab.org/documents/frus1945v02d128

Doubtless, readers can think of 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

Our proposal for a new measure of US diplomatic priorities is 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. Diplomatic importance (or status) is “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).[30] Existing literature (Singer and Small 1966; Small and Singer 1973) has used diplomatic representation as an indicator for countries’ diplomatic importance and/or status in international politics, largely based on the number of diplomatic missions countries received in particular.[31]

However, these datasets were not created to measure any micro-level variation in inter-state relations. Indeed, the seminal dataset on diplomatic importance and recognition was not intended to measure “the importance of one particular state to another particular state” and its authors emphasize that “no inferences should be---or can be---made about China's importance to the United States from the absence of their respective embassies in Peking and Washington” (Small and Singer 1973, 580).[32]

Moreover, the US has refused to recognize some hostile countries for political reasons despite their strategic and diplomatic importance because US diplomatic representation itself can serve as a signal of US intentions to engage with the host country. Thus, the US had been constrained or unwilling to signal such intention regarding adversaries or parties in negotiation with the US, such as North Korea or Cuba.

[30] Numerous works have noted the importance of status in international politics (e.g. O’Neill 2001; Lake 2014).

[31] Renshon (2017) also measures status by deriving a network centrality measure with community detection algorithms from the diplomatic exchange data (Bayer 2006). Relatedly, Arias and Smith (2018) measure the perceived prestige of US diplomatic posts abroad by ranking the host countries by size and wealth while noting that “[r]anking the prestige of postings is a non-trivial task” (96).

[32] Recent scholarly efforts to improve the data on diplomatic representation include measures of diplomatic “focus” that reflects “the level of diplomatic attention being devoted to a given Hosting Country by a given Guesting Country” by assessing whether each diplomatic mission is physically located in the host country and whether it is appointed to multiple countries (Moyer, Bohl, and Turner 2016b, 5).

Relatedly, political importance of a country to the US has also been measured by its membership, regular participation, and voting behavior in the UN General Assembly and Permanent Membership in the UN Security Council (Lebovic and Saunders 2016). While useful for a longer time horizon, this measure arguably reflects *de jure* diplomatic importance rather than *de facto* importance to US foreign policy in the 1970s. In particular, relatively smaller states that belatedly joined the UN but became salient issues on the US strategic or domestic political front---such as North Korea, Vietnam, or the German Democratic Republic (East Germany)---are not considered “important” in the early 1970s by this measure.[33] The measure also does not the capture the variation in diplomatic importance of countries that are involved in crises/events but have been members of the UN since the early days, such as Iran or Afghanistan.[34] Finally, there is a potential endogeneity to the measure. A country can vote with the US because its leaders believe that doing so will bring in more foreign aid or World Bank loans. Rather than reflecting political importance to the US, voting similarity could reflect the importance of the US to that country.

Our measure complements these efforts, measuring countries’ relative diplomatic importance to the US in the 1970s. Our granular data allows us to 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 for US diplomatic priorities is constructed at the country-year level, but can be even more fine-grained depending on a researcher’s needs. We measure them by counting the number of cables tagged for each country per calendar year. Appendix C includes summary statistics of the country-year- and country-level measures for countries existent during the period 1973-79, including only non-US countries and excluding countries that disappeared before 1973 or countries that emerged after 1979 (“contemporary” countries) (Correlates of War Project 2017).[35] Our measure exists for 1,040 country-years and 156 countries from 1973--1979. 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. It can be tagged in as many as 24,856 cables in a given year and 144,726 in total.

Below are 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

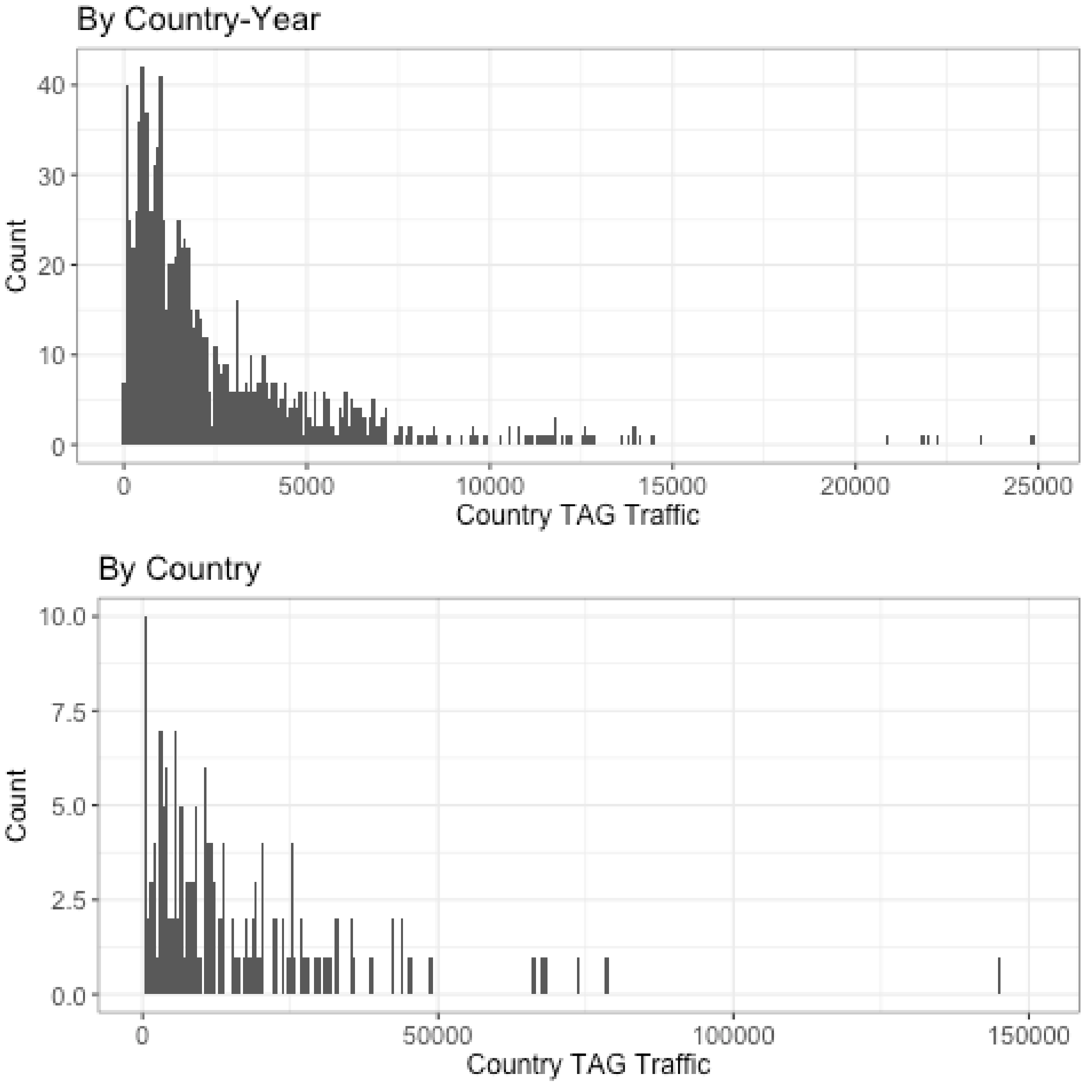
[33] The United Nations recognized the the German Democratic Republic in 1973, Vietnam in 1977, and North Korea in 1991. Note that the US vetoed South Vietnam and North Vietnam’s application to join the UN in 1975 and the reunified Vietnam’s application in 1976.

[34] Iran and Afghanistan joined the United Nations in 1945 and 1946, respectively.

[35] Appendix C also includes summary statistics of country TAG traffic for all country-years and countries, including the US and countries that existed before 1973 or after 1979.

countries produce heavy TAG traffic. 85.86% of country-years were tagged in fewer than 5,000 cables, 10.48% between 5,000 and 10,000 cables, and only 3.65% in more than 10,000 cables. Similarly, 76.92% of countries were tagged in fewer than 25,000 cables, 21.15% between 25,000 and 75,000 cables, and only 1.92% in more than 75,000 cables.

FIGURE 4: Country TAG Traffic at Country-Year and Country Levels



Note that country TAG traffic is different from cable traffic. 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 in 1974. The corresponding figures for 1975 through 1979 were: 23,404, 24,856, 21,836, 22,244, and 21,978 tagged with the Soviet Union and only 21,187, 22,552, 12,492, 14,355, and 13,640 cables sent to or received from Moscow or Leningrad.[36] This implies that the Soviet Union was frequently

[36] See Appendix D for a full comparison.

discussed in cables that were not directly sent or received by the US embassy or Consulate General in the USSR. This also indicates that country tag traffic is a more meaningful measure that better captures the importance of the Soviet Union to US foreign policy than the number of cables to or from the Soviet Union. Moreover, variations in year-to-year country tag traffic with a particular country could indicate major events that occurred in the country. For example, while Iran was always one of the countries with the most tags during this period, the number of mentions spiked in 1979 with the Iranian Revolution and the Hostage Crisis.

Our data fills a gap 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 in 1973-74 or North Vietnam/Vietnam in 1974-79, or Rhodesia in 1978 (Arias and Smith 2018; Baggott Carter 2018a; Lebovic 2018; Moyer, Bohl, and Turner 2016a). However, these countries frequently appeared in our measure for diplomatic importance using the cables in the CFPF corpora. The German Democratic Republic was tagged in 10,551 diplomatic cables in 1974 and 3,521 cables in 1973.[37] North Vietnam/Vietnam was tagged in 8,384 cables in 1979; 4,830 in 1978; 4,148 in 1977; 1,907 in 1976; 3,054 in 1975; and 2,028 in 1974. Rhodesia was featured in 3,903 cables in 1978.[38]

5.2. Validation

In this section, we informally validate our measure of country-year's diplomatic importance. We do a "sanity check" of the measure, checking the countries and country-years with the highest and lowest volumes of country TAG traffic.[39] Our measure shows both a static component to the most important countries but also some dynamic fluctuation with political events.

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 German Democratic Republic were in the first and third positions, respectively. Four developed country allies or economic competitors are in the top 10: United Kingdom (3rd), Japan (4th), France (8th), and Canada (10th). The remaining countries reflected regional powers: Israel (5th), Egypt (6th) and Iran (9th).

[37] Our TAGS are drawn from US State Department metadata released after the documents were reviewed for declassification. The reviewers frequently added a GE tag for cables that covered West Germany. GE was the code for East Germany, so total traffic for East Germany is overcounted. We have manually extracted TAGS for West Germany from the original TAGS field available in the raw text of the body.

[38] The State Department has separate TAGS for North Vietnam (Democratic Republic of Vietnam) and Vietnam (Socialist Republic of Vietnam), reunited with South Vietnam (the Republic of Vietnam) in 1976. We are in the process of updating SQL tables for these countries.

[39] We exclude the US and non-contemporary countries---countries that did not exist anytime in the period of 1973-79---in this section for compatibility with other datasets. See Appendix E for the lists of 40 country-years and 40 countries with the highest and lowest country TAG traffic.

Not surprisingly, these countries also dominate the list of 20 country-years with the largest amount of TAG traffic. All 6 years with the Soviet Union are included as well as 4 years with the United Kingdom and the German Democratic Republic. But the other country-years reflect major political events. Iran's traffic in 1979 is represented, the year of the Iranian Revolution and hostage crisis. Traffic concerning Israel and Egypt in 1979, which witnessed the negotiation and implementation of the Camp David Accords, also appears in the top 20---as does traffic about the Republic of Vietnam in 1975. Rather than reflecting only static importance, our measure also shows changing priorities.

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. Furthermore, the measure gives more weight to countries' geopolitical importance than their size, measured by population. Despite their population, neither China (mean population of approximately 941 million in 1973-79) nor India (616 million) cracked our top 10 list of country traffic. Other very populous countries that did not receive corresponding diplomatic traffic include: Indonesia (133 million); Brazil (110 million); Bangladesh (80 million); Pakistan (73 million); Nigeria (64 million) according to the National Material Capabilities dataset version 5.0 (Singer, Bremer, and Stuckey 1972). Moreover, our measure correctly ranks the Soviet Union first in importance instead of third by population, the United Kingdom second instead of 12th, the German Democratic Republic third instead of 35th, Israel fifth instead of 95th, Egypt 6th instead of 19th, France 7th instead of 14th, Iran 10th instead of 23rd.[40]

6. Discussion

In this paper, we introduce CHRONOS, a new database of over 3 million documents about diplomacy and foreign policy from the US, UK, and Brazil. We provide corpora of previously classified internal government documents and their metadata and full text that spans the period of 1861-2013.

Our datasets provide many new opportunities for IR scholars on diplomacy due to their "deep" nature. They capture events, actions, information, and preferences of both adversarial and non-adversarial diplomacy, private and public diplomacy, and diplomacy at multiple micro-levels of analysis---individuals, offices, agencies, and departments.

In particular, 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 CHRONOS Database also includes an online platform with an API and a website from which scholars can view the text and associated statistics online and download datasets.

[40] See Appendix F for a full comparison of top 20 countries' ranks in country TAG traffic and population.

We also demonstrate how our data can be used in research on diplomacy by constructing a measure of US diplomatic priorities from the documents in the Central Foreign Policy Files, a collection of over 3.2 million U.S. diplomatic documents from 1973 to 1979. This provides a better measure of the relative importance of countries like North Vietnam, China (PRC), and East Germany (GDR), than a binary measure of whether they were diplomatically recognized by the US.

Our data comes with some caveats, as with any new collection. As discussed, full text of some documents in the CFPF collection is unavailable, either because it was not preserved by the State Department or it has not been released to researchers. The collection also does not include documents previously classified as “Top Secret.” However, we provide metadata of all non-Top Secret documents, regardless of their withdrawal status.

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. The preprocessing stage involved manual examination where we sampled and compared cleaned and original data to improve the quality. We also pre-processed the documents with additional Python scripts for TAG corrections. For example, we discovered that certain country TAGS were rarely stored as metadata but exist in the raw versions of the cables. For important cases, such as the the Federal Republic of Germany (West Germany) and the Republic of Vietnam (South Vietnam), we have manually extracted it from the raw documents and stored it as corrected 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.

CHRONOS shows how NLP tools can aid in the organization, exploration, and analysis of large corpora of once-secret records for social science research. Without such methods, political scientists, sociologists, economists, and historians would be hard-pressed to cope with millions of email, texts, and other media that constitute the source base for future research. Since declassified government documents are already in the public domain, they present an opportunity to develop new research standards and methods for future work using textual data, whether in the US or worldwide.

References

Acharya, Avidit, and Kristopher W. Ramsay. 2013. “The Calculus of the Security Dilemma.”

Quarterly Journal of Political Science 8: 183–203.

Allen, David, and Matthew Connelly. 2016. “Diplomatic History after the Big Bang: Using Computational Methods to Explore the Infinite Archive.” In *Explaining the History of American Foreign Relations*, edited by Frank Costigliola and Michael J.Editors Hogan, 3rd ed., 74–101. Cambridge University Press.

Arel-Bundock, Vincent, James Atkinson, and Rachel Augustine Potter. 2015. “The Limits of Foreign Aid Diplomacy: How Bureaucratic Design Shapes Aid Distribution.” *International Studies Quarterly* 59 (3): 544–56.

Arias, Eric, and Alastair Smith. 2018. “Tenure, Promotion and Performance: The Career Path of US Ambassadors.” *The Review of International Organizations* 13 (1): 77–103.

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

———. 2018b. “Diversionary Cheap Talk: Unemployment and US Foreign Policy Rhetoric, 1945-2010.” http://www.erinbcarter.org/documents/diversionUS.pdf.

Bayer, Resat. 2006. *Diplomatic Exchange Data Set*. http://correlatesofwar.org.

Bayne, Nicholas, and Stephen Woolcock. 2016. “The New Economic Diplomacy.” *Decision-Making and Negotiation in International Economic Relations, Aldershot, Ashgate Publishing Limited*.

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.

Boehme, Franziska. 2018. “Exit, Voice and Loyalty: State Rhetoric about the International Criminal Court.” *The International Journal of Human Rights* 22 (3): 420–45.

Bueno de Mesquita, Bruce, James D. Morrow, and Ethan Zorick. 1997. “Capabilities, Perception, and Escalation.” *American Political Science Review* 91 (1): 15–27.

Bull, Hedley. 1977. *The Anarchical Society: A Study of Order in World Politics*. Second. New York: Columbia University Press.

Carnegie, Allison. 2014. “States Held Hostage: Political Hold-Up Problems and the Effects of International Institutions.” *American Political Science Review* 108 (1): 54–70.

———. 2015. *Power Plays: How International Institutions Reshape Coercive Diplomacy*. Cambridge University Press.

Carson, Austin. 2016. “Facing Off and Saving Face: Covert Intervention and Escalation Management in the Korean War.” *International Organization* 70 (1): 103–31.

———. 2018. *Secret Wars: Covert Conflict in International Politics*. Princeton University Press.

Carson, Austin, and Keren Yarhi-Milo. 2017. “Covert Communication: The Intelligibility and Credibility of Signaling in Secret.” *Security Studies* 26 (1): 124–156.

Cogburn, D. L., and A. Wozniak. 2013. “Computationally Intensive Content Analysis of Public Diplomacy Data: Understanding the Public Remarks of US Secretaries of State, 1997-2011.” In *2013 46th Hawaii International Conference on System Sciences*, 1269–78.

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

Crawford, Vincent P., and Joel Sobel. 1982. “Strategic Information Transmission.” *Econometrica* 50 (6): 1431–1451.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. “Reputation and Status as Motives for War.” *Annual Review of Political Science* 17 (1): 371–393.

Dietrich, Simone. 2016. “Donor Political Economies and the Pursuit of Aid Effectiveness.” *International Organization* 70 (1): 65–102.

D’Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt. 2014. “Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines.” *Political Analysis* 22 (2): 224–42.

Drezner, Daniel W. 2000. “Ideas, Bureaucratic Politics, and the Crafting of Foreign Policy.” *American Journal of Political Science* 44 (4): 733–49.

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.

Farrell, Joseph, and Robert Gibbons. 1989. “Cheap Talk Can Matter in Bargaining.” *Journal of Economic Theory* 48 (1): 221 – 237.

Fearon, James. 2013. “Fighting Rather than Bargaining.” https://web.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/frtb6.pdf.

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

———. 1995. “Rationalist Explanations for War.” *International Organization* 49 (3): 379–414.

———. 1997. “Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs.” *Journal of Conflict Resolution* 41 (1): 68–90.

Fey, Mark, and Kristopher W. Ramsay. 2010. “When Is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation.” *World Politics* 62: 529–560.

Fuchs, Andreas, and Nils-Hendrik Klann. 2013. “Paying a Visit: The Dalai Lama Effect on International Trade.” *Journal of International Economics* 91 (1): 164–177.

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.

Goldsmith, Benjamin E., and Yusaku Horiuchi. 2009. “Spinning the Globe? U.S. Public Diplomacy and Foreign Public Opinion.” *The Journal of Politics* 71 (3): 863–75.

Gray, Julia, and Philip BK Potter. 2017. “Diplomacy and the Settlement of International Disputes.” https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3012639.

Guisinger, Alexandra, and Elizabeth N. Saunders. 2017. “Mapping the Boundaries of Elite Cues: How Elites Shape Mass Opinion across International Issues.” *International Studies Quarterly* 61 (2): 425–441.

Guisinger, Alexandra, and Alastair Smith. 2002. “Honest Threats: The Interaction of Reputation and Political Institutions in International Crises.” *Journal of Conflict Resolution* 46 (2): 175–200.

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. 2015. *Emotional Diplomacy: Official Emotion on the International Stage*. 1st ed. Cornell University Press.

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.

Hollibaugh, Gary E. 2015. “The Political Determinants of Ambassadorial Appointments.” *Presidential Studies Quarterly* 45 (3): 445–66.

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

———. 2018. *Face-to-Face Diplomacy: Social Neuroscience and International Relations*. Cambridge, UK: Cambridge University Press.

Jervis, Robert. 1970. *The Logic of Images in International Relations.* Princeton, NJ: Princeton University Press.

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

———. 2001. “Signaling and Perception: Drawing Inferences and Projecting Images.” In *Political Psychology.*, edited by Kristen Renwick Monroe, 293–312. Mahwah, NJ: Lawrence Erlbaum Associates.

———. 2011. *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War*. Cornell University Press.

———. 2017a. *How Statesmen Think: The Psychology of International Politics*. Princeton University Press Princeton.

———. 2017b. *Perception and Misperception in International Politics*. 2nd ed. Princeton, NJ: Princeton University Press.

Kastner, Scott L., and Phillip C. Saunders. 2011. “Is China a Status Quo or Revisionist State? Leadership Travel as an Empirical Indicator of Foreign Policy Priorities.” *International Studies Quarterly* 56 (1): 163–77.

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.

Kentikelenis, Alexander, and Erik Voeten. 2018. “Exit, Voice, and Loyalty towards Liberal International Institutions: Evidence from United Nations Speeches 1970-2017.” https://www.internationalpoliticaleconomysociety.org/sites/default/files/paper-uploads/2018-11-01-14_52_56-ev42@georgetown.edu.pdf.

Kilby, Christopher. 2011. “What Determines the Size of Aid Projects?” *Expanding Our Understanding of Aid with a New Generation in Development Finance Information* 39 (11): 1981–94.

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

———. 2013b. “Dependent Diplomacy: Signaling, Strategy, and Prestige in the Diplomatic Network.” *International Studies Quarterly* 58 (2): 247–59.

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

Kydd, Andrew. 2003. “Which Side Are You On?: Bias, Credibility, and Mediation.” *American Journal of Political Science* 47 (4): 597–611.

Lake, David A. 2014. “Authority, Status, and the End of the American Century.” In *Status in World Politics*, edited by T. V. Paul, Deborah Larson, and William Wohlforth, 246–72. Cambridge University Press.

Lebovic, James H. 2018. “Security First?: The Traveling U.S. Secretary of State in a Second Presidential Term.” *Presidential Studies Quarterly* 48 (2): 292–317.

Lebovic, James H., and Elizabeth N. Saunders. 2016. “The Diplomatic Core: The Determinants of High-Level US Diplomatic Visits, 1946–2010.” *International Studies Quarterly* 60 (1): 107–23.

Leventoglu, Bahar, and Ahmer Tarar. 2005. “Prenegotiation Public Commitment in Domestic and International Bargaining.” *American Political Science Review* 99 (03): 419–433.

———. 2008. “Does Private Information Lead to Delay or War in Crisis Bargaining?” *International Studies Quarterly* 52 (3): 533–553.

Li, Xiaoting. 2015. “Dealing with the Ambivalent Dragon: Can Engagement Moderate China’s Strategic Competition with America?” *International Interactions* 41 (3): 480–508.

Lindsey, David. 2017. “Diplomacy Through Agents.” *International Studies Quarterly* 61 (3): 544–56.

Lindsey, David, and William Hobbs. 2015. “Presidential Effort and International Outcomes: Evidence for an Executive Bottleneck.” *The Journal of Politics* 77 (4): 1089–1102.

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

Maliniak, Daniel, and Michael Plouffe. 2011. “A Network Approach to the Formation of Diplomatic Ties.” SSRN Scholarly Paper ID 1900231. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=1900231.

McManus, Roseanne W. 2014. “Fighting Words: The Effectiveness of Statements of Resolve in International Conflict.” *Journal of Peace Research* 51 (6): 726–40.

McManus, Roseanne W. 2017a. *Statements of Resolve: Achieving Coercive Credibility in International Conflict*. Cambridge: Cambridge University Press.

———. 2017b. “The Impact of Context on the Ability of Leaders to Signal Resolve.” *International Interactions* 43 (3): 453–79.

———. 2018. “Making It Personal: The Role of Leader-Specific Signals in Extended Deterrence.” *The Journal of Politics* 80 (3): 982–95.

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

Moyer, Jonathan D., David K. B Bohl, and Sara Turner. 2016a. “Diplomatic Representation Data Set | Pardee Center for International Futures.” Diplometrics: Diplomatic Representation [Data File]. 2016. https://pardee.du.edu/diplomatic-representation-data-set.

———. 2016b. “DIPLOMETRICS: Diplomatic Representation Data Codebook (Embassy Codebook) Version 3.16.16.” 2016. http://diplodash.pardee.du.edu/downloads/Diplometrics_Diplomatic_Exchange_Codebook_V3.16.16.pdf.

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

O’Neill, B. 2001. *Honor, Symbols, and War*. Political Science. University of Michigan Press.

Palmer, Glenn, Vito D’Orazio, Michael Kenwick, and Matthew Lane. 2015. “The MID4 Dataset, 2002–2010: Procedures, Coding Rules and Description.” *Conflict Management and Peace Science* 32 (2): 222–42.

Pekkanen, Saadia M., Mireya Solís, and Saori N. Katada. 2007. “Trading Gains for Control: International Trade Forums and Japanese Economic Diplomacy.” *International Studies Quarterly* 51 (4): 945–70.

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.

Potter, Philip BK. 2013. “Electoral Margins and American Foreign Policy.” *International Studies Quarterly* 57 (3): 505–518.

Powell, Robert. 2004. “Bargaining and Learning While Fighting.” *American Journal of Political Science* 48 (2): 344–361.

Putnam, Robert. 1988. “Diplomacy and Domestic Politics: The Logic of Two-Level Games.” *International Organization* 42 (3): 427–460.

Ramsay, Kristopher W. 2011. ““Cheap Talk" Diplomacy, Voluntary Negotiations, and Variable Bargaining Power.” *International Studies Quarterly* 55: 1003–1023.

Rathbun, Brian C. 2014. *Diplomacy’s Value: Creating Security in 1920s Europe and the Contemporary Middle East*. Ithaca, NY: Cornell University Press.

Regan, Patrick M., Richard W. Frank, and Aysegul Aydin. 2009. “Diplomatic Interventions and Civil War: A New Dataset.” *Journal of Peace Research* 46 (1): 135–46.

Renshon, Jonathan. 2008. “Stability and Change in Belief Systems: The Operational Code of George W. Bush.” *Journal of Conflict Resolution* 52 (6): 820–49.

———. 2009. “When Public Statements Reveal Private Beliefs: Assessing Operational Codes at a Distance.” *Political Psychology* 30 (4): 649–61.

———. 2017. *Fighting for Status: Hierarchy and Conflict in World Politics*. Princeton University Press.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. “Leader Influence and Reputation Formation in World Politics.” *American Journal of Political Science* 62 (2): 325–39.

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.

———. 2005. *Deterrence by Diplomacy*. Princeton, NJ: Princeton University Press.

Saunders, Elizabeth N. 2015. “War and the Inner Circle: Democratic Elites and the Politics of Using Force.” *Security Studies* 24 (3): 466–501.

———. 2017. “No Substitute for Experience: Presidents, Advisers, and Information in Group Decision Making.” *International Organization* 71 (S1): S219–S247.

Schelling, Thomas C. 1966. *Arms and Influence.* New Haven: Yale University Press.

———. 1980. *The Strategy of Conflict.* Cambridge: Harvard University Press.

Singer, J. David, Stuart Bremer, and John Stuckey. 1972. “Capability Distribution, Uncertainty, and Major Power War.” In *Peace, War, and Numbers.*, edited by Bruce M. Russett, 19–48. Beverly Hills, CA: Sage.

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

Slantchev, Branislav. 2003a. “The Power to Hurt: Costly Conflict with Completely Informed States.” *American Political Science Review* 97 (1): 123–133.

———. 2003b. “The Principle of Convergence in Wartime Negotiations.” *American Political Science Review* 97 (4): 621–632.

Slantchev, Branislav L. 2006. “Politicians, the Media, and Domestic Audience Costs.” *International Studies Quarterly* 50 (2): 445–477.

Small, Melvin. 1977. “Doing Diplomatic History by the Numbers: A Rejoinder.” *Journal of Conflict Resolution* 21 (1): 23–34.

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.

Smith, Alastair. 1998. “International Crises and Domestic Politics.” *American Political Science Review* 92 (3): 623–638.

Tomz, Michael. 2012. *Reputation and International Cooperation: Sovereign Debt across Three Centuries*. Princeton University Press.

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

Trager, Robert F. 2016. “Diplomacy of War and Peace.” *Annual Review of Political Science* 19 (1): 205–28.

———. 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.

Watson, Adam. 1984. *Diplomacy, the Dialogue Between States*. E. Methuen.

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

Xierali, Imam M., and Lin Liu. 2006. “The Effect of Power and Space on Foreign Diplomatic Presence in the United States: A Spatial Modeling Approach.” *Geographic Information Sciences* 12 (2): 53–63.

Yarhi-Milo, Keren. 2013. “In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries.” *International Security* 38 (1): 7–51.

———. 2014. *Knowing The Adversary: Leaders, Intelligence Organizations, and Assessments of Intentions in International Relations*. Princeton, NJ: Princeton University Press.

———. 2018. *Who Fights for Reputation? The Psychology of Leaders in International Conflict*. Princeton, NJ: Princeton University Press.

Yarhi-Milo, Keren, and Alex Weisiger. 2015. “Revisiting Reputation: How Do Past Actions Matter in International Politics.” *International Organization* 69 (2): 473–95.

Appendix A: Details on the CFPF Corpus

In Section 4.4., we introduced the Central Foreign Policy Files corpus. Here we provide additional details about the corpus. TABLE 2 shows the number of documents in the corpus with non-missing values by variable (field). Most of the key metadata variables have non-missing values.

TABLE 2: Number of Cables with Non-Missing Values by Variable

Variable	Number of Documents with Non-Missing Values
collection	3214293
id	3214293
body	2654414
date	3214293
classification	2654414
subject	2876678
from_field	3214094
to_field	3213050
concepts	3063262
office	2654414
handling	2654414
type	3214293

TABLE 3 displays the number of documents in the corpus by year. It shows that the documents in our corpus are spread out evenly throughout the period.

TABLE 3: Number of Cables by Year

Year	Number of Cables	Relative Frequency
1973	179253	5.58%
1974	442301	13.76%
1975	531102	16.52%
1976	554864	17.26%
1977	474671	14.77%
1978	500577	15.57%
1979	531525	16.54%
Total	3214293	-

Below we show the number of documents in the corpus by classification level. The majority, namely 57.2%, of the documents are those previously unclassified. The corpus also includes some documents previously classified “Secret,” however.

TABLE 4: Number of Cables by Classification Level

Classification	Number of Documents	Relative Frequency
Secret	127332	4.8%
Confidential	494823	18.64%
Unclassified	1518305	57.2%
Limited Official Use	513769	19.36%
Total	2654229	-

Appendix B: Details on the FRUS Corpus

In Section 4.5., we discussed the Foreign Relations of the United States corpus. Here we provide additional details about the corpus.

TABLE 5 shows number of documents with non-missing values by variable/field. The lack of missing values for the variable “body” implies that all documents’ full text is available. However, some documents are missing dates and many classification.

TABLE 5: Number of FRUS Documents with Non-Missing Values by Variable

Variable	Number of Documents with Non-Missing Values
collection	209046
id	209046
body	209046
date	186279
classification	52580
volume_id	209046
chapt_title	178050
title	209034
p_from	97657
p_to	51797
source	59028

TABLE 6 below displays the number of documents in the FRUS corpus by classification level. While the corpus includes many documents with missing classification, it includes documents previously classified as “Top Secret” and “Secret.”

TABLE 6: Number of FRUS Documents by Classification Level

Classification	Number of Documents	Relative Frequency
	156466	74.85%
Confidential	13512	6.46%
Secret	29937	14.32%
Top Secret	9131	4.37%
Total	209046	-

Appendix C: Summary Statistics of Country TAG Traffic

In Section 5.1., we discussed some descriptive statistics of country TAG traffic for non-US countries that existed anytime during the period of 1973-79, a measure we derived from the documents in the CFPF corpus. Here we show full summary statistics of country TAG traffic.

Below we show summary statistics of country TAG traffic for contemporary non-US countries by country-year (TABLE 7) and by country (TABLE 8).

TABLE 7: Summary Statistics by Country-Year (Only Contemporary Non-US Countries)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Year	1,040	1,976.1	2.0	1,973	1,974	1,978	1,979
COW Codes of Countries	1,040	460.3	247.3	20	253.8	663	990
Country TAG Traffic	1,040	2,545.4	3,019.4	21	678.2	3,394.5	24,856

TABLE 8: Summary Statistics by Country (Only Contemporary Non-US Countries)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
COW Codes of Countries	156	459.6	253.6	20	233.8	663.8	990
Country TAG Traffic	156	17,036.6	19,338.0	277	4,643	22,983.2	144,726

TABLES 9 and 10 display summary statistics of country TAG traffic of country-years and countries, including the US and all countries that have ever existed, such as Azerbaijan, Bavaria, Belarus, Kazakhstan, Saxony, Tajikistan, and Tuscany.

TABLE 9: Summary Statistics by Country-Year (Incl. Non-Contemporary Countries and the US)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Year	1,519	1,976.0	2.0	1,973	1,974	1,978	1,979
COW Codes of Countries	1,519	460.0	256.6	2	271	670	990
Country TAG Traffic	1,519	2,220.0	7,652.6	0	33.5	2,224.5	138,438

TABLE 10: Summary Statistics by Country (Incl. Non-Contemporary Countries and the US)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
COW Codes of Countries	217	460.0	257.1	2	271	670	990
Country TAG Traffic	217	15,540.2	50,372.7	0	277	18,121	705,142

Appendix D: Comparison of Country TAG Traffic with Cable Traffic

In Section 5.1., we compared country TAG traffic with cable traffic in the case of the USSR. Here we provide further details about the comparison. Below we show the number of cables tagged with the USSR, the number of cables sent by or to the US embassy in Moscow, and the number of cables sent or to the US Consulate General in Leningrad.

TABLE 11: Country TAG Traffic vs. Cable Traffic (USSR Case)

Year	Number of Cables Tagged with the USSR	Number of Cables Sent by/to the US Embassy in Moscow	Number of Cables Sent by/to< the US Consulate General in Leningrad
1,973	9,532	10,149	209
1,974	20,876	17,246	999
1,975	23,404	20,217	1,388
1,976	24,856	21,598	1,377
1,977	21,836	11,867	844
1,978	22,244	13,616	1,080
1,979	21,978	13,196	1,999

Appendix E: Highs and Lows of Country TAG Traffic

In Section 5.2., we informally validated our measure of diplomatic importance, country TAG traffic, by examining the extreme values. Here we show the highs and lows of country TAG traffic by country-year (TABLES 12 and 13) and country (TABLES 14 and 15) among non-US contemporary countries.

TABLE 12 below displays 20 country-years with the largest volume of country TAG traffic.

TABLE 12: Non-US Country-Years with Highest Country TAG Traffic

Year	Tagged Country	Number of Cables	Relative Frequency
1976	Soviet Union	24856	0.74%
1975	Soviet Union	23404	0.7%
1978	Soviet Union	22244	0.66%
1979	Soviet Union	21978	0.66%
1977	Soviet Union	21836	0.65%
1974	Soviet Union	20876	0.62%
1979	Iran	14433	0.43%
1977	United Kingdom	14145	0.42%
1979	Israel	13974	0.42%
1978	Israel	13918	0.42%
1976	German Democratic Republic	13775	0.41%
1977	German Democratic Republic	13606	0.41%
1976	United Kingdom	12885	0.38%
1979	Egypt	12764	0.38%
1978	German Democratic Republic	12733	0.38%
1978	United Kingdom	12630	0.38%
1979	United Kingdom	12605	0.38%
1975	Republic of Vietnam	12551	0.37%
1975	German Democratic Republic	12228	0.36%
1975	Japan	12087	0.36%

TABLE 13 below displays 20 country-years with the smallest volume of country TAG traffic.

TABLE 13: Non-US Country-Years with Lowest Country TAG Traffic

Year	Tagged Country	Number of Cables	Relative Frequency
1977	Mongolia	75	0%
1979	Maldives	75	0%
1978	Equatorial Guinea	72	0%
1979	Bhutan	68	0%
1977	Sao Tome and Principe	67	0%
1975	Mongolia	66	0%
1974	Bhutan	63	0%
1977	Equatorial Guinea	57	0%
1973	Albania	55	0%
1975	Maldives	55	0%
1978	Mongolia	50	0%
1979	Mongolia	48	0%
1973	Equatorial Guinea	45	0%
1973	Bhutan	35	0%
1975	Bhutan	31	0%
1977	Bhutan	31	0%
1976	Bhutan	28	0%
1973	Maldives	27	0%
1973	Congo	23	0%
1978	Bhutan	21	0%

TABLE 14 below lists 20 countries with the largest volume of country TAG traffic.

TABLE 14: Non-US Countries Most Frequently Tagged in Cables

Country	Number of Cables	Relative Frequency
Soviet Union	144726	4.3%
United Kingdom	78832	2.34%
German Democratic Republic	78192	2.33%
Japan	73518	2.19%
Israel	68113	2.03%
Egypt	67582	2.01%
France	65907	1.96%
Mexico	48875	1.45%
Canada	48519	1.44%
Iran	45385	1.35%
Italy	44763	1.33%
China	43965	1.31%
India	43688	1.3%
Thailand	42668	1.27%
German Federal Republic	42379	1.26%
South Korea	38899	1.16%
Turkey	38411	1.14%
South Africa	35767	1.06%
Philippines	35227	1.05%
Poland	35157	1.05%

TABLE 15 below lists 20 countries with the smallest volume of country TAG traffic.

TABLE 15: Non-US Countries Least Frequently Tagged in Cables

Country	Number of Cables	Relative Frequency
Gambia	2401	0%
Congo	2082	0%
Seychelles	1897	0%
Guinea-Bissau	1786	0%
Yemen People's Republic	1772	0%
Grenada	1745	0%
Albania	1571	0%
Cape Verde	1332	0%
Djibouti	1188	0%
Equatorial Guinea	950	0%
Samoa	665	0%
Dominica	621	0%
Maldives	577	0%
Comoros	577	0%
Mongolia	553	0%
Sao Tome and Principe	541	0%
Solomon Islands	521	0%
St. Lucia	496	0%
St. Vincent and the Grenadines	354	0%
Bhutan	277	0%

Appendix F: Comparison of Country TAG Traffic with Population

In Section 5.2., we compared countries’ rank in country TAG traffic and mean population. Below we provide a full list of top 20 countries’ ranks in country TAG traffic and ranks in mean population in 1973-79.

TABLE 16: Rank in Country TAG Traffic vs. Rank in Population

Top 20 Countries in Country TAG Traffic	Rank in Country TAG Traffic	Rank in Mean Population
Soviet Union	1	3
United Kingdom	2	12
German Democratic Republic	3	35
Japan	4	5
Israel	5	95
Egypt	6	19
France	7	14
Mexico	8	11
Canada	9	30
Iran	10	23
Italy	11	13
China	12	1
India	13	2
Thailand	14	16
German Federal Republic	15	10
South Korea	16	21
Turkey	17	17
South Africa	18	27
Philippines	19	15
Poland	20	22