4. Indexing the Empire

Sarah Harrison

WHAT IS PLUSD?

"The Public Library of US Diplomacy," or "PlusD," is a very large and constantly expanding collection of internal documents from the US Department of State, published by WikiLeaks in a searchable archive. The library began in 2010 and at the time of writing contains 2,325,961 individual documents made up of about 2 billion words, spread over three collections of cables: *Cablegate*, the Kissinger Cables, and the Carter Cables. The State Department is the foreign affairs department of the US government and oversees the embassies and consulates of the United States all over the world. Each embassy or consulate corresponds with the State Department in Washington, DC, by sending daily telegram reports, or "cables," between them, using a special electronic communications system.

PlusD contains within it the WikiLeaks publication known as *Cablegate*: the collection of State Department cables published by WikiLeaks in 2010 and 2011. *Cablegate* itself consists of 251,287 cables, accounting for 261,276,536 words in total. If printed out in a standard-sized font, *Cablegate* alone would form a single line over 6,000 kilometers long—the distance to the center of the Earth. The cables are an average of 1,039 words long, revealing detailed internal information about the operation of 274 US embassies and consulates, and their activity within their host country.

WikiLeaks specializes in publishing, curating, and ensuring easy access to full online archives of information that has been censored or suppressed, or is likely to be lost. An understanding of our historical record enables self-determination; publishing and ensuring easy access to full archives, rather than just individual documents, is central to preserving this historical record. Since publishing *Cablegate*, WikiLeaks has continued to work to make PlusD the most complete online archive of US Department of State documents, adding to the library each year with newly available cables and other documents from the State Department communications system. It can be accessed through a set of specially developed search interfaces at **https://wikileaks.org/plusd.**

HOW TO READ THE CABLES

Probably the first question for anyone researching a serious topic in the cables is: Where do I start? Experience has shown that the answer is: Don't start by searching for specific things.

Because of what it is—the archive of a foreign ministry—PlusD is a rich repository of information on countries, major international and domestic figures, political parties, events, policies, processes, trends, and developments. There is a natural temptation to "mine" PlusD: to think of particularly notable topics from the news, then to dig a narrow shaft down through the huge amount of information to find the cables where only that topic is mentioned, treating those cables as an authority on the subject. Much of the early reportage on the cables in the mainstream media was done this way.

This approach is not necessarily a misuse of the cables—there is plenty of information in them that is responsive to this kind of query. But, in general, reading the cables this way will result in a superficial understanding of them, and of the subject of the research. Because there is so much information in PlusD, it is easy to find information that confirms your biases as a researcher, with the result that you will see only what you want to see. If these pitfalls are to be avoided, you must think about your reading strategy from the outset.

First, try to get a good feel for what kind of material you are handling. The cables were not written to provide instant information on a range of discrete topics to a general readership. Instead, they are the means the State Department uses to communicate with itself—the by-product of the daily operation of embassies all over the world. The original way they were read was as updates on preceding documents, in a continuous succession.

This gives the documents features which must be kept in mind if they are to be understood properly. First, the people who wrote the cables are diplomats: they are specialists in communicating with each other, and this means they assume a lot of prior knowledge. Often, to understand a cable it is necessary to understand what is *not* said in the cable but left implicit; in order to do this it is necessary to read other cables to get a more general picture.

Second, the cables are episodic. Each of them is part of a succession of cables over time, reporting how—to the best of the knowledge of the authors—situations are unfolding in the country in question. Without appreciating the dynamic nature of the subject-matter of the cables, and the fact that the authors are often working with incomplete knowledge of that subject-matter, it is easy to miss out on the rich historical insight the documents offer.

The obvious remedy to this is to read widely around the topic you wish to research, and to become as familiar as possible with the documentary context of your topic. If your research focuses on a particular country, a good way to do this is to take the highest classification level for that country (which will be a small subset) and read all of the cables in it, chronologically. If you are reading about a particular event, make sure to define a period covering that event, and read all of the cables from the same embassy within that time period. If your topic centers on a few keywords, such as a particular figure, or a hot issue within the cables, such as "extraordinary rendition" or "genetically modified organisms," make sure that, once you have located cables responsive to your search queries, you also scan the cables for related issues and topics—reading widely within those searches too. Try also to read other cables sent around the same time. Understanding that the US government actively lobbies foreign governments to encourage deregulation of GMOs is, in isolation, a narrow insight; understanding how this policy evolves over time, and how it interrelates with Washington's other long-term diplomatic goals in that country, furnishes a broader understanding, and it will inform any reading about similar topics.

Make sure at all times to maintain a critical distance from the documents. When the *New York Times* offered an overview of the cables, it remarked that the cables broadly confirmed the dominant view of the US as a benevolent superpower, upholding American values and advocating for human rights abroad. This is unsurprising, if you consider that the *New York Times* shares the same ideology of US exceptionalism that is compulsory in the State Department. As the output of the US diplomatic community, the cables will reflect the biases and ideology of the US government and establishment, and its aspirations in the wider world. It is important always to be on the lookout for how ideology is shaping the content: the euphemisms and clichés, and the way in which contentious issues are hidden in plain sight, or left out entirely.

For example, the concept that US oil corporations are entitled to extract and export the natural resources of Venezuela and Libya would be too brazen a concept for US diplomats to endorse explicitly. Hence, it is reframed in the cables as "resource nationalism," to make it seem as if it is a bad thing when the government of Venezuela decides that the natural resources of Venezuela should principally benefit Venezuelans. It is only by reading widely that it is possible to understand the full implications of a concept like "resource nationalism," and how it is involved with US foreign policy, and thereby to be able to read it against the grain.

Bias in the US diplomatic corps is not monolithic, either. Policies and orthodoxies change over time, and in the cables you can see diplomats amending their language as perspectives change—choosing to observe institutional taboos, or to pander to new policy obsessions coming from Washington. Different administrations—the Bush and Obama presidencies are very well covered in the documents—also usher in different priorities and emphases.

Besides bias and ideology, it is important to remember that specific information in a cable can be inaccurate. Sometimes reports will reflect incomplete information; sometimes diplomats will engage in speculation that turns out, in hindsight, to have been unfounded; and, sometimes, what they report will be simply wrong. But these cables are still important, as they provide a genuine record of what information was being sent to or from an embassy at that time.

KNOWING YOUR WAY AROUND A CABLE

As with any specialized document, there are some things it is important to understand about the cables in order to read them properly.

First, metadata. This is what the cable tells you about itself: its unique reference number, its date, where it comes from, where it was sent, what kind of subjects it touches on, which other cables it references, its classification level and handling restrictions, and other specific information about each cable. In the PlusD archive we have processed the metadata of each document, presenting it in a special box at the top of each cable. For each class of metadata, you can click on the metadata field to see more information. You can, for example, click on the classification for more information about the classification scheme and how it applies to the cable you are reading. You can also choose to view the "raw header"—the metadata as it was before we processed it. All these fields are individually searchable in the PlusD search interface.

Each cable also has a text title—for example, "EGYPT: GAZA ROUND-UP: DECEMBER 31"—and a date that is exact to the minute—for example "2009 January 1, 02:45 (Thursday)." Each cable also has an official reference ID given to it by the State Department—a unique reference number that is meant to refer only to that cable. This is typically of the form "09CAIRO1." The first two digits, "09," denote the year the cable was sent: 2009; the middle of the cable ID indicates the origin (the US embassy in Cairo); and the final digits indicate the sequential number of the cable that year (in this case, it is the first cable of the year 2009: "1"). In some cases, the State Department's reference ID system breaks down and cables are given duplicate names. In PlusD, we have created a canonical ID which ensures that all cables have a unique identifier across all datasets, rectifying any mistaken duplications by the State Department. This canonical ID is created by taking the original document ID and adding a "_" at the end, followed by WikiLeaks' annotation for different datasets: *Cablegate* = a, the Kissinger Cables = b, and so on. For document IDs that were duplicated in the original datasets, we number each duplicate—for example, 1976WARSAW05657_b2 is the second document with that State Department ID in the Kissinger Cables.

Like most government agencies, the State Department uses classification to restrict access to information on a "need to know" basis. Cables are assigned a classification level depending on how sensitive they are, and only people with the corresponding "clearance" can read those cables. The higher the classification, the smaller the set of people who are allowed official access. Some

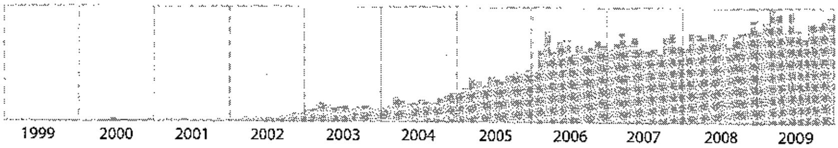

Figure 1: Frequency of US Department of State Cables between 1999 and 2009

cables also have handling restrictions, such as "NOFORN," meaning the cable cannot be shown to any non-US nationals, or "FOR OFFICIAL USE ONLY." The *Cablegate* set, for example, does not contain any cables with the highest classification level ("TOP SECRET"), but does contain cables with every classification beneath this. There are 4,330 cables classified as SECRET//NOFORN. There are 11,322 classified as SECRET. Some 4,678 cables are classified as CONFIDENTIAL//NOFORN, and 97,070 as CONFIDENTIAL. Furthermore, 58,095 cables are marked UNCLASSIFIED//FOR OFFICIAL USE ONLY, and 75,792 are marked UNCLASSIFIED. Most embassies will have some cables within each classification.

Classification level does not necessarily equate with significance or "newsworthiness." Even if a cable is marked "SECRET//NOFORN," this does not mean that the information contained within it will be more sensational or interesting for your purposes than information contained in a document with a lower classification level. The reason is normally given for why a particular cable has been given its classification, and reasons often include the fact that a cable deals with a past operation, or communications security, or contains information that would be embarrassing, either to the US government or to the host government, if disclosed. The content of these cables can sometimes seem quite pedestrian, and the classification might have been given to it simply because it was formally required. But there will often be other cables from that period, possibly at a lower classification level—or even unclassified—that contain important comments by a senior diplomat shedding light on a US perspective on a national issue, or in aggregate disclosing an historically significant or important insight. It is therefore important not to assume that the lowest-level security classification denotes the lowest level of political significance.

Sometimes the classification schemes are not rigorously adhered to by the diplomats. Particularly with some of the smaller embassies, there is either no facility for TOP SECRET ciphers, or it is used so infrequently that the opportunity costs involved with using it encourage diplomats to take shortcuts. The result is that information that by all rights should be highly classified is sometimes given in cables with a lower classification level.

Each cable is normally marked with one or more "TAGS." "TAGS" refers to "Traffic Analysis by Geography and Subject"—a categorization system implemented by the State Department in 1973 to group cables more effectively according to their subject matter and geographical relevance. TAGS range from country codes ("GR" for Germany, "IZ" for Iraq, and so on) to organization codes ("AEC" for Atomic Energy Commission, "DOD" for Department of Defense, and so on), program codes ("KCIP" for Critical Infrastructure Protection, "KMPI" for Middle East Partnership Initiative, and so on), and subject codes ("PROP" for Political Affairs—Propaganda and Psychological Operations, "PREL" for Political Affairs: External Political Relations, and so on). WikiLeaks has researched the acronyms and expanded all of them so that they can be read without having to consult a key. You can now click each tag to see other cables to which it is attached. A full list of the TAGS acronyms can be found at http://wikileaks.org/plusd/about-ta.

Watch out for "reftel," which is the internal citation system for the cables. If the cable you are reading uses the word "reftel," this is a reference to a previous cable in which the topic is expanded upon. Normally, reftels are listed at the top of the body of the cable. If the reftel is one of the cables contained in the PlusD dataset, it should be hyperlinked, and you should be able to read the reftel simply by clicking on it. Following the thread of reftels is often a useful way of getting a full impression of the events or topic dealt with across a succession of cables.

The body of each cable is more straightforward. Cables are divided into numbered paragraphs. The cables will typically take the form of a report on a meeting or discussion that has taken place within the diplomatic premises, or as part of the official business of the diplomatic mission. Commentary is often included in brackets around the main report. The first paragraph is normally a summary of the cable. The rest of the cable goes into more detail. Cables are normally signed off by one of the senior officials within the diplomatic mission: the ambassador, the chargé d'affaires, the political officer, the consul, or another official who is made responsible for a specific set of duties. But this is a formality, and an individual's signature does not mean that that person actually read the cable. Many cables begin with a brief note explaining which official within the embassy or consulate classified the cable, and for what reason.

Reading the cables will turn up some unfamiliar terms and acronyms, such as "POLOFF" (political officer) and "SIPDIS" (SIPRNet Distribution). WikiLeaks has assembled a comprehensive explanatory database of known terms at https://wikileaks.org/plusd/tags to aid researchers in understanding the information in context.

HOW TO ACCESS PLUSD

The best way to read State Department cables published by WikiLeaks is through our special search interface at https://wikileaks.org/plusd. It is impossible to provide a comprehensive summary of the research tools available on PlusD here, but full instructions are available on the website. Researchers can choose which collection to search (*Cablegate*, Kissinger Cables, and so on, or several at once) and can choose to search within specific date ranges, specific geographical regions, specific embassies or consulates, and within specific classification levels, among many other fields and tools available for precise searching and research. For example, PlusD will generate a graph from any search term, showing you the frequency of the occurrence of that search term within the whole of the PlusD database.

PlusD allows users to limit their searches with reference to fields from the cable metadata. Some of the most interesting of these are the lesser-known ones. The office field refers to the particular office or bureau within the State Department that the cable was to or from, and whether it was regarding an "action" by the State Department ("to"), or originated in the State Department ("from"). This field allows researchers either to narrow their search results for a certain field (for example, narrowing a search to documents to or from the Committee on Oceans and Atmosphere for someone researching DOS communications on fisheries), or to gain a better understanding of how the State Department is dealing with a certain topic (for example, if researching a topic that is discussed in cables copied specifically to the Bureau of Intelligence and Research, it allows for a more nuanced understanding of how the United States views this topic).

There are some fields searchable in PlusD that the *Cablegate* collection does not record—for example, Handling Restrictions. The Handling Restrictions field provides for a more detailed understanding of who was and was not allowed to see each document, over and above the classification level a person would need to hold, by stating the allowed range of distribution—for example, Exclusive Distribution Only (EXDIS), which indicates "extremely limited dissemination." To prepare this field for PlusD, WikiLeaks not only extracted the field from the metadata of the document, but searched in the raw data of the cable for the word "Cherokee," which appears 2,208 times in the Kissinger Cables and 1,263 times in the Carter Cables, and extracted this as one of the searchable handling restrictions possible. The word "Cherokee" is reserved for messages involving the Secretary of State and senior White House officials only. The term originated during the 1960s, when Secretary of State Dean Rusk named it after Cherokee County, Georgia, where he was born. Due to the limited distribution of cables carrying this handling restriction, it is a rare and important addition to the possible entries in this field, only specifically searchable in PlusD.

The PlusD text- and field-search interface facilitates searching and search refinement across seventeen different fields, including additional explanations of what abbreviated entries in each field mean. Other interfaces available in PlusD to search the archive include mapping occurrences of certain words over time and browsing frequencies of TAGS used in the documents. This variety of tools allows all types of researchers to access the full PlusD archive for searches both broad and narrow.

WikiLeaks has been publishing classified or otherwise suppressed documents and archives since 2006. These are not just cables, but include a diverse range of documents—from internal military reports and government documents to suppressed studies and investigative work, internet filter lists, and internal bank documents. A dedicated global search engine for every single document WikiLeaks has published can be found at https://**search.wikileaks.org**. A guide to using the global search engine can be found at https://**search.wikileaks.org/info**.

Since late 2012, we have included a tool that allows readers to highlight the parts of the cable they find most interesting and link other internet users directly to that material. The highlighter can be found at the bottom-right of the screen on the PlusD reader.

A full overview of how PlusD was prepared by WikiLeaks, providing insight into the structure of the cables, can be found at http://**wikileaks.org/plusd/about**.

PUBLISHING PLUSD

The first collection in PlusD was *Cablegate*, which was originally published in 2010 as part of a partnership of international newspapers and media organizations globally, coordinated by WikiLeaks. We designed and implemented a system that allowed us to coordinate a publication schedule between over a hundred global mainstream media partners. Whenever the media partners were to publish a story, they would enter into this system the cables they were going to use in their story, so that WikiLeaks would publish the cable at the same time. These partnerships ran for almost a year, after which—because WikiLeaks holds fast to the principle that full archives should be published—we ensured that every single cable was published in full. All of them can now be read online.

Through the partnerships, WikiLeaks' media partners were under a memorandum of understanding (MOU) to publish the full text of the cables (initially redactions were permitted in a few very specific circumstances outlined in the MOU) when their story went live, but this did not always happen. The redactions, according to the MOU, were to be made only if a specified and identifiable individual would be at real risk of death or punishment with no judicial process. However, the press often abused this agreement, and in many cases redacted for entirely different reasons—for example, political bias. In addition, many media published only extracts from selected cables, or did not publish the cable at all. Since WikiLeaks published the full unredacted archive, the public has had unhindered access to the record. This has resulted in the exposure of journalistic error and bias, and has enabled the global readership of *Cablegate* to become active participants in the interrogation of our historical record.

There are nearly a quarter of a million cables in *Cablegate*, from as early as 1966, although there is a thinner distribution of cables over the earlier decades than there is for recent years. The bulk of the cables in *Cablegate* are from the State Department under the George W. Bush and Barack Obama administrations, thus relating to the decade beginning around the year 2000. The most recent cables in the collection are from early 2010. PlusD also contains collections of cables that originally became available through US Government declassification procedures. In order to ensure that these cables could not be unpublished or reclassified by the government (a common occurrence), and in order to make the documents more visible and searchable, WikiLeaks incorporated them into PlusD in two collections, depending on the date of their release by the US Government: at the time of writing, this meant that there were three individual collections of cables in the PlusD archive. In early 2013, we published the "Kissinger Files"—that is, 1,707,500 diplomatic documents originating between the years 1973 and 1976, the Kissinger years; and in April 2014, we published the "Carter Cables"—367,174 diplomatic cables from the year 1977.

The creation of PlusD involved complex data journalism and archival processes, which included manually processing each cable and correcting spelling errors introduced by the State Department in indexing information. Thanks to our work, cables tagged by the State Department as, for example, "Brasil" and "Brazil" are now indexed as referring to the same country in PlusD. In some cases, our journalistic partners were able to discover twice as many cables in response to a single search term as a consequence of our work. PlusD is consequently the most comprehensive and powerful database of US diplomatic cables in existence. As more State Department cables become available, whether through declassification or the brave actions of whistleblowers, we will continue to grow the PlusD database.

One of WikiLeaks' principles is to provide the public with the resources to inform itself, and this means ensuring the data are presented in a manner that ensures easy interaction and research for all. Some of our hardest work goes toward adding value to datasets and making our publications more accessible and usable. This involves researching the structure of the data, designing and implementing search engines, optimizing metadata, and adding a large number of features to make the data easier to navigate and explore for researchers, journalists, human rights groups, historians, students, and others.

Over the years, we have improved our search interface and sought to contextualize the cables, making them more accessible and navigable. Our efforts have been reflected in the continued use of our publications by the media and the public alike. PlusD continues to be an invaluable resource for investigative journalists looking for context and background for developing stories. Every day, new stories are published in mainstream news publications that explicitly reference the *Cablegate* archive. There is not a significant geopolitical event in the world that cannot be illuminated with material published by WikiLeaks. We expect PlusD to continue to yield crucial historical insight long into the future.

USE OUR WORK

WikiLeaks undertakes to publish information of diplomatic, ethical, or historical significance that has been censored, suppressed, or is under threat of being lost to history. This information is frequently available only through the actions of courageous individuals within secretive organizations: whistleblowers. Commensurate with the risks taken by such individuals, WikiLeaks undertakes to protect our journalistic sources with the best, most advanced techniques available. We promise our sources that we will publish in such a way as to produce the maximum impact possible. We promise to publish in full, and that once something has been published it will never be unpublished.

Our work is dedicated to making sure history belongs to everyone, not just to elite organizations and their counterparts in the news industry. By publishing source documents, WikiLeaks helps to ensure accountability on the part of not only those with executive power, but also the media. If you use our publications in your research and writing, make sure to link to the source document, publicize your discoveries widely, and demand of every other news organization that it does not hold back or suppress the common history of humanity.

Donations to WikiLeaks are welcome, at https://**wikiLeaks.org/ donate.**