Making Open-Access Data More Accessible: Learnings from Harnessing GDELT 2.0 Queries for the General Audience

Elizabeth Gooch[a], Eric Eckstrand[b]

[a]*Defense Resources Management Institute, Naval Postgraduate School*
[b]*Data Science Analytics Group, Naval Postgraduate School*

Abstract

This article presents the conceptual and technical framework for our creation of a browser-based access tool that draws on a subset of the massive open-access Global Database Events Locations and Time (GDELT) 2.0. GDELT is an enormous dataset on the global media updating every 15 minutes since March 2015. In theory, the database is a gold-mine of coded data on events and related articles for interested researchers to quantify trends in the world over time and space. However, the size and access limit most potential users ability to utilizer GDELT. We have constructed a user-friendly interface that enables point-and-click queries of a Chinese activities in the media and the subsequent coverage of those events.

Keywords:

We started off this project with an assumption. We think that if open-access data was more technically approachable for social science researchers, journalists, and other interested parties lacking coding skills, then more analyses would exist to benefit the world. This is a particularly compelling quandry at this time when data is growing exponentially in quantity and topic. In this essay, we outline the steps taken to make a large and, arguably important, a portion of Global Database Events Locations and Time (GDELT) 2.0, the broadest collection of global new articles available today, more accessible to a general audience (GDELT, 2015, 2020c). We developed a

Email addresses: elizabeth.gooch@nps.edu (Elizabeth Gooch), eric.eckstrand@nps.edu (Eric Eckstrand)

tool called GDELT Browser-Based Access Tool, or GBBAT to non-technical users to query a GDELT. We believe that the ideas we used to creating GBBAT are generalizable to other open-access big data.

The rest of the paper is organized as follows. In the next section, we preview the major take-aways from our experience developing GBBAT. Section 2 provides a brief introduction to GDELT data. In section 3, we review the conceptual framework for creating GBBAT and in section 4 we detail the technical consideration and quandries we faced during the app's development. Section 6 concludes.

1. Preview of Results

Our analysis is broken into conceptual and a technical components. The learning for our readers also can be divided along these line.

2. Data: GDELT Details

GDELT curates the data via algorithm-based variables based on text analysis that provided deeper insight into the emotional tone of articles as well as the topics addressed. While GDELT has undertaken steps to make the dataset more accessible through a speedy interface via the Google BigQuery cloud platform, the GUI still requires the user to query in SQL language. Additionally, there are other data quality issues embedded within the data collection structure that would corrupt analyses for user without advanced data management skills.

The 2.0 version of GDELT offers three tables: event, event mentions, and global knowledge graph (GKG). Together, these tables provide characteristics of global events and all of the reporting on those events. The events table is small in size and content, including world events at the two actor observations. Two actors is an important characteristic of how global news dataset usually classify events. One event, to us, as members of world society, will likely be included in a data table like events as more than two event observations depending on the pairwise interaction of participating actors.

The event mention table draws on the events observations from the events table, capturing each news article published about an event. While the events table produced by GDELT 2.0 has analogs such as.., the event mentions table is a unique offering from GDELT. No other news collection organization records all subsequent articles published about a global event. Typically,

only the first article to report the event is captured in other global news data products. The GKG table is also a special contribution of GDELT to the global research community but remains the hardest to access, at more than 1 TB of information over the past five years.

Linking across the three tables of GDELT 2.0 is the feature that is most useful about the GDELT. But the computing power required to filter and merge data from these three tables is large. For example, an interesting question can be asked of the global media and GDELT can provide a snapshot of an answer. For example, someone may want to know what portion of Chinese activities located in West African nations are agricultural versus economic?

3. Conceptual Framework

GDELT has a problem that plagues many open-access databases. The database contains a lot of information that analysts would be interested in wielding but the data is inaccessible to those without technical computers science skills such as data wrangling, data management, data storage, and programming skills. Open-access data or other forms of non-proprietary data is becoming an increasingly prevalent phenomenon due to greater storage power and increased data collection opportunities. Many of these open-access datasets are very large. Application of these data products for a better understanding of the world, however, requires analysis. Analysts with subject matter expertise, however, often do not also have the data wrangling skills to wield the open-access datasets. It is an either-or problem. Either a researcher has subject matter expertise on data subject or a researcher has computer science expertise on how to deftly manipulate the open-access data.

Here is one way to think about what we are doing: If an interdisciplinary team is a fundamental answer to drawing useful analysis from open-access big data, then, in a sense, we design a working team that is not contemporanous but instead is a multi-stage process. First, the political economy expertise narrows the topical focus of GDELT. Next, the computer scientists begin the process by focusing on data and data relations that would interest analysts and develop a user interface. Finally, interested researchers can conduct analyses on the available data to address subject matter issues through the interface.

3.1. Subsetting the Dataset

A fundamental characteristic of the open-access data that we are dealing with is that it is enormous. The computer scientist team knows that large-scale data management requires storage options that are costly. Creating a useful subset of the full dataset should be the first step for the team.

We chose a portion of the open-access data that will be most attractive to analysts. To illustrate this consideration, suppose as a computer scientist team, you had access to a hypothetical dataset known to house all data in the whole world. What analysts would be attracted to your project? Many analysts would be overwhelmed. And the ones that did attempt to use the interface you created for this broad dataset would choose to look into narrow questions. At this step, we suggest starting off with an attractive narrow yet insightful subset of the overall open-access dataset.

For GDELT 2.0, we chose to focus on China in the media, specifically, articles addressing events in which the country of origin of the actor in an event in from China. Since 2013, Chinese economic and diplomatic overseas activities have surged with the national Belt and Road strategy. Not only is China aboard is a key topic in recent global news, but also the years of data that GDELT 2.0 offers, 2015-2020, coincided well with the onset of the Belt and Road initiative. Moreover, Chinese influence abroad is a central component of the U.S. National Security Strategy since 2018 and our initial funding is from the Department of Defense interests.

To further refine the subset of the data that we made available to non-technical users, we exclude variables that we deemed unnecessary because of repetitiveness or those made obsolete by our assumptions in the construction of the subset.

3.2. Addressing Analytical Problems

GDELT 2.0 has one serious analytical problem in that the origin of global media changes over time. This means, for example, that sources contributing to global media in 2020 are larger in number than in 2016. What is problematic about this characteristic of GDELT, is not just that the global media is expanding, but that researchers do not why or how it is expanding and the trends in expansion may affect the conclusions they draw from their analyses. We addressed this issue by confining reporters to a stable list.

The data science teams should also highlight the important variables for analysis quality. For GDELT 2.0 the confidence level for which the event mention is categorized is important. Confidence level runs from 0 to 100,

where 0 means that GDELT categorized any article that seems to link to the event at all as linked to the event. While a confidence level of 100 means that GDELT is certain that the article links to the event that it is coded with. Therefore, the subset of data at the 0 confidence level is the largest containing all events at higher confidence levels.

Along the same lines, as a data science team, it is important to omit problematic observations. Here is where the team is making another judgment call that should be made clear to the non-technical user. Our experience with GDELT offers an illustrative example. The average length of many articles was one sentence. This short article is not what an analyst expects. The analyst expects to be tallying news articles. In GDELT 2.0, for example, one of the top reporters in the full dataset is iHeart radio. This is not the source that an analyst on China is interested in citing for their research. Only data-savvy analysts would be able to identify iHeart radio observations out at the event mention-level for removal, let alone if they had already merged with the GKG, then the iHeart radio computing power would be eating up the precious resources of the filter and merge function.

3.3. Creating Value-added

In GDELT 2.0, the media source identification is not of primary concern for GDELT, but is if researchers wanted to draw conclusions such as what is the average opinion from the Ghanian media on the in-country Chinese activities? To surmise the source origin, the URL has to be manually, visually sorted by the backend data scientist. Recording the range of sources on which the data subset is based is certified by the web application created and the process and outcomes are presented in a transparent nature on the homepage for the project. This is also a step of cleaning the data but we cleaned out observations that could hurt the analysts' research goals.

We foresee the manual data cleaning of press sources as one of the major bottlenecks. However, limiting the sources addresses the problem highlighted in 3.2 and makes our data product more useable.

4. Technical Description and Design Decisions

In this section we provide a technical description of the browser-based exploration tool we developed, which sits on top of GDELT data. In some cases, we decided it was preferable to choose one technical implementation

over multiple, competing implementations. In these cases, we will provide our rationale for the choices we made.

4.1. Data Source

The GDELT project records news events in a persistent storage structure. In the event that power is removed from the system, recorded events are not lost. Maintaining data in a persistent state, also facilitates its transfer from one party to another. Numerous mechanisms exist for data persistence, but the GDELT project uses two primary methods, which are publically accessible. First, the data is stored in comma-separated value (CSV) files. Each file is composed of GDELT data spanning a 15-minute interval. Recording started in February of 2015 and has proceeded until the present day. An up-to-date listing of all CSV files produced can be downloaded from the project website (GDELT, 2020b). Each row in this listing contains a URL to an individual zip-compressed CSV file for a specific 15 minute interval and GDELT table (**gkg**, **export**, or **mentions**). So, to obtain data for an entire 24-hour period, for example, you must download 96 separate files. A file naming convention is used, which makes it convenient to select a specific 15-minute interval. A file name starts with a timestamp (YYYYMMDDHHHHmm) and is followed by the GDELT table, and is terminated with the "csv.zip" extension. If you wanted all of the GDELT Global Knowledge Graph (GKG) data for the 15 minutes starting on September 8th, 2020 at 1515 (UTC), then you would access the URL *http://data.gdeltproject.org/gdeltv2/20200908151500.gkg.CSV.zip*.

The GDELT 2.0 Event Database has also been persisted in the Google Cloud Platform (GCP). The entire up-to-date database of records can be accessed via the Google BigQuery interface (GDELT, 2020a). There are a number of advantages associated with this persistence and access scheme. First, the GDELT database, and all of its constituent tables, are hosted in GCP, which can be readily queried using standard SQL syntax. In contrast, the CSV-download API described above allows you to download 15-intervals of data, but it does not provide a mechanism to conduct table-wide exploration in situ. For example, if you wanted to know the number of records that exist in the GDELT Events table starting from the time of table creation, using the CSV-download API you would first need to download all of the 15-minute intervals to a local workstation. Next, all of the individual files would need to be unzipped. Assuming that the resulting data did not overwhelm the local storage capacity, a custom program to count the number of records

in each file would need to be developed and executed. In contrast, using the BigQuery cloud-based access scheme, this same task could be performed simply by issuing an SQL count query.

The convenience of using the BigQuery interface to explore and analyze GDELT data comes at a monetary cost. The most significant cost, in this regard, is associated with performing on-demand queries. An individual query costs $ 5.00 per terabyte Google (2020). GDELT tables range in size with the Events table consisting of hundreds of gigabytes of records and the GKG table consisting of tens of terabytes of records. These tables are ever-growing as new records are being inserted continuously by the GDELT maintainers. Of note, BigQuery users do not pay for storage costs associated with GDELT project.

Considering our budgetary and technology infrastructure constraints, we decided to download a subset of the GDELT Events and Event Mentions tables using the CSV-download API. After unzipping each of the individual files, we inserted the records into MySQL database tables, one table for the Events data and one table for the Event Mentions data. The Events table consisted of 436 million records starting from February 2015 and ending in May 2019 for a total of 160 gigabytes. The Event Mentions table consisted of 1.4 billion records covering the same date range for a total of 287 gigabytes. By utilizing a MySQL solution, we were able to expose a convenient SQL interface that our analysts and applications could utilize without needing a GCP account or having to perform cumbersome CSV download operations.

4.2. Data Preparation, Cleaning, and Filtering

While an SQL interface to GDELT data is an improvement over a CSV download scheme, not every analyst is proficient in SQL. We have developed a web browser based tool, which allows the user to formulate GDELT queries through a click-button GUI interface. There is no need for the analyst to formulate their queries in SQL. The interface is intuitive, so that it is easy to filter GDELT along many dimensions. Additionally, the tool provides data visualization for enhanced understanding. Our tool also offers a CSV data download capability for those users that would like to export query results for follow-on analysis, perhaps in a different tool. However, before GDELT data is presented to the end-user through the GBBAT GUI, the source data is transformed in a number of ways. The *GDELT Browser-Based Access Tool* (GBBAT) tool does not provide unfettered access to our entire MySQL database of GDELT data. We take a number of steps to clean and filter

the source dataset in order to prepare it for end-user queries, which are guaranteed to return in a timely manner.

Our data preparation steps are based on a number of assumptions, which coincide with the anticipated interests of GBBAT users. First, we assume that users are only interested in China-related events. We implemented this assumption by removing any record in the Events table where the Actor1CountryCode attribute or the Actor2CountryCode attribute are not *China*. This resulted in a 97.5 percent reduction in the number of records in the Events table. Second, we have deemed a subset of news sources to be relevant. The exact list of news sources can be found in the Appendix. We implemented this assumption by removing any record in the Event Mentions table where the MentionSourceName attribute value is not within the news-source-relevant set. This resulted in a 98.3 percent reduction in the number of records in the Event Mentions table.

The next data preparation step that we undertook, prior to exposing the data to the end-user for querying, was appending Event table information to the records in the Event Mentions table. For example, we appended the Actor1Name , Actor2Name, Actor1CountryCode, and Actor2CountryCode from records in the filtered Events table to records in the filtered Event Mentions table, where both records share a common GLOBALEVENTID attribute value. This operation can be thought of as augmenting the Event Mentions records with information from the Events table. The resulting table consisted of approximately 3 million records. A number of unintersting attributes were dropped from this table. The overall size of the final table was 400 megabytes. The entirety of the merging operation is performed by a standard SQL INNER JOIN. The exact SQL sytax for the entire data preparation process can be found in the Appendix.

The above-mentioned data preparation steps are only performed once and the final 400 MB table is kept in our MySQL database. User queries generated through the GBBAT utility are only issued against this table, and not against any of the source or intermediate tables generated in the data preparation procedure. This is done to keep the GBBAT utility responsive. Due to the size of the source Events and Event Mentions tables, allowing the user to perform arbitrary SQL operations against these tables could result in significantly delayed responses.

4.3. Data Queries and Presentation

Only a subset of SQL query operations are made available through the GBBAT click-button interface. Given the final table described in the previous section, we allow the user to further filter the table based on Actor1CountryCode and Actor2CountryCode, the EventCode, the ActionGeoCountryCode, and the PressOrigin. For example, if the end-user wanted all Event Mention records that took place in *Bangladesh*, where the press origin was *China*, the other actor was the *United States*, and the records were tagged with the CAMEO code of *Engage in diplomatic cooperation*, then the user could select each of these options through drop-down menus. Next, the user could simply press a button to download each of the records that matched this query as a CSV file.

We also provide a number of built-in visualizations based on the user-submitted query. For example, we plot the tone of each of the records in the Event Mentions table over time on the user-filtered results.

The following figures highlight the basic functionality of the GBBAT tool.

5. Technical Recommendations

In order for GBBAT queries against the MySQL table to be performant, we found it necessary to create table indexes for the filterable columns (Actor1CountryCode, Actor2CountryCode, EventCode, ActionGeoCountryCode, and PressOrigin). Additionally, even thought the final table is only created once, the data preparation steps were sped up by indexing certain key columns of the source tables.

One of the downsides of creating a static snapshot of the GDELT dataset is that the end-user is limited to a specific period of time. Records that occur after this period of time are not available. It is incumbent upon the application maintainer to obtain new records and update the final table. If the application maintainer has limited availability to perform these update operations, then the end-user may not be able to obtain the latest records. It is possible that the update procedure could be automated by using a differnent data store. The GDELT dataset in BigQuery has the most up-to-date records present. We intend on investigating switching our data store from a static MySQL-based snapshot to a dynamic, up-to-date BigQuery solution in future work, along with any monetary cost differences.

6. Concluding Remarks

Non-technical researchers can breach technical hurdles themselves by cultivating interdisciplinary teams of subject matter experts and data scientists. However, not every analyst has this option. For one, it requires the development of long-term, professional relationships. In academia, interdisciplinary research may be stifled by the incentive structure of department promotion which rewards publication within the field. In journalism, the role of data scientists is uncommon. Not to mention, the inaccessibility of research team creation at earlier stages of schooling, i.e. science fair projects, masters' thesis, etc. Like our team, which is supported by the U.S. Department of Defense, interdisciplinary teams may be more possible that the government and private-sector research are able to create research environments that sustain interdisciplinary teams.

Our team is uncommon in that is it comprised of data scientists, economists, and political scientists. It is likely that our mish-mash of experience propelled us towards developing a web interface in which the data scientists have already completed the wrangling and at the second stage, the analyst (at various skill levels of data management) can explore the open-access data for which they would otherwise have been unable to tap into due to their poorly developed programming ability and access to data storage frameworks.

References

GDELT, 2015. The GDELT event database data format codebook v2.0. Tech. rep., GDELT PROJECT.
URL http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf

GDELT, 2020a. GDELT Big Query Interface.

GDELT, 2020b. GDELT Master CSV list.
URL http://data.gdeltproject.org/gdeltv2/masterfilelist.txt

GDELT, 2020c. Our global world in realtime.
URL http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

Google, 2020. Big Query Pricing.
URL https://cloud.google.com/bigquery/pricing