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

Keywords:

1. Introduction

We are writing this practice piece to outline the steps that we took to make a large and, arguably important, a portion of Global Database Events Locations and Time (GDELT) 2.0 more accessible to a general audience (GDELT, 2015, 2020). We believe that the steps we took with GDELT 2.0 can be generalized to other open-access big data.

Here, we make our first assumption for this project. We think that if the open-access data was more approachable for social science researchers, journalists, and other interested parties without coding skills, then, through their analyses the world may benefit from the uniting of technical and subject matter knowledge. The relative lack of publications, particularly social science publication based on GDELT 2.0 in Google Scholar may be evidence to support our assumption.

The problem is that many open-access data contain 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

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

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.

This is the case with GDELT 2.0. GDELT offers the broadest collection of global new articles available today. Additionally, GDELT curates the data by offering 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. Moreover, there are other data quality issues embedded within the data collection structure that would corrupt analyses for user without advanced data management skills.

Non-technical researchers could breach these 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.

As one of the rare interdisciplinary teams made up of data scientists, economists, and political scientists, our 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.

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

but temporarily spaced out - a two-stage process. The computer scientists begin the process by focusing on data and data relations that would interest analysts and develop a user interface. Afterward, the subject experts or subject matter interested researchers can conduct analyses on the available data through the interface.

2. Data: GDELT details

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

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.

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. In subsequent sections, we will refer to the tool that we developed as the GDELT Browser-Based Access Tool, or GBBAT.

4.1. Data Source

In the process of recording real-world events, as GDELT does, the information that is collected is often kept in a persistent structure. In the

event that power is removed from the system holding recorded events, existing information is not lost. Keeping data in a persisted state, such as this, can also facilitate the transfer of information from one party to another. There are numerous persistence schemes available for storing data. GDELT data are persisted in two, primary ways. First, the data is available in 15-minute intervals, starting in February of 2015 and proceeding until the present day, from the GDELT project website. The master list of all the files produced by the project can be downloaded from the following URL: http://data.gdeltproject.org/gdeltv2/masterfilelist.txt. Each 15-minute interval is formatted in a comma-separated value (CSV) file, which is then compressed. To obtain data for an entire day, for example, you can download 96 separate, zip-compressed CSV files. A naming convention for files is used, which makes it easy to select a specific 15-minute interval. The name of a files starts with a timestamp and is followed by the desired GDELT table, and is terminated with the ".csv.zip" extension. A specific 15-minute interval could be downloaded by accessing the URL in this format: http://data.gdeltproject.org/gdeltv2/YYYYMMDDHHHHMM.[gkg—mentions—export].csv.zip.

The method described above for identifying and transferring GDELT records to your local workstation is convenient, but it is not conducive to follow-on data tasks, such as exploration and visualization. Fortunately, the GDELT 2.0 Event Database has also been persisted in the Google Cloud Platform (GCP). More specifically, the entire database of records can be accessed via the Google BigQuery tool. 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 web API provided by the GDELT organization allows you to download 15-intervals of data, but it does not provide a mechanism to conduct table-wide exploration. For example, if you wanted to know how many records existed in the mentions table, starting from the time of table creation, then under the web API persistence scheme, 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. Under the cloud-based access scheme, the same task could be performed simply by issuing a SQL count query against the Google BigQuery interface.

4.2. Data Preparation, Cleaning, and Filtering

While the BigQuery interface to a cloud-based copy of the GDELT dataset is a great improvement over a flat-file CSV scheme, where the GDELT data is chunked into 15-minute intervals and persisted to a local workstation, not every analyst is proficient in SQL. We have developed a web-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. For those users that would like to export query results for follow-on analysis in another tool, the web-interface offers a CSV data download capability. Finally, a token-based API is exposed for those who do not wish to use the GUI. Some of the details of this architecture are presented in the next section.

The GBBAT tool does not provide unfettered access to the entire GDELT dataset. We take a number of steps to clean and filter the source dataset in order to prepare it for end-user queries. 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 records. In terms of the events table, we interpret this to mean that either the Actor1CountryCode or the Actor2CountryCode must be China. The events table is ever-growing, but at the time of this writing, China-related events, as we have defined them, account for approximately X percent of the records, which represents a significant reduction in the number of total events. Second, we have deemed a subset of news sources to be relevant to our users. The exact list of news sources can be found in the Appendix. In terms of the mentions table, this cleaning operation reduces the number of records considerably. Once again, the mentions table is ever-growing, but at the time of this writing, this cleaning operation reduced the size of the mentions table by about X percent.

The next step in the data preparation phase that we undertake, prior to exposing the data to the end-user for querying, is appending event information to each of the records in the mentions table. For example, we append the event's Actor1Name , Actor2Name, Actor1CountryCode, and Actor2CountryCode to each mention record in our cleaned mentions table. We also drop a number of columns from the cleaned mentions table, which we deem uninteresting from the perspective of our user-base. This entire operation is performed by a standard SQL JOIN operation. The exact SQL queries

for the entire data preparation process can be found in the Appendix. The resulting table is kept in GCP under our account separate from the GDELT project tables, and is made available to the GBBAT application for specific user queries.

The filter and join procedure described above is non-trivial, which take around X minutes to complete and about X dollars. Given our time and budget constraints, we cannot perform the data preparation steps every time there is an update to GDELT (every 15 minutes). Instead, we re-perform the data preparation steps once, at the beginning of each day.

4.3. Data Queries and Presentation

As discussed, the GBBAT query tool does not provide the complete functionality available to the SQL query language. Instead, we provide a narrow subset of query options through a click-button interface. Given the table described in the previous section, we allow the user to further filter the table based on Actor1CountryCode and Actor2CountryCode, the EventCode, or CAMEO code, the ActionGeoCountryCode, and the PressOrigin. For example, if the end-user wanted event mentions that took place in Bangladesh, where the press origin was China, and the other actor was the United States, and were tagged with the CAMEO code of 05, then the user could select each of these options through GUI-based drop-down menus. Next, the end-used could simply press a button to download the results as a CSV file. The end-user can then perform further analyses using a spreadsheet tool of their choosing, for example.

We also provide a number of built-in visualizations based on the user-submitted query. For example, we plot the tone of the event mentions over time on the user-filtered results.

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

5. Technical Recommendations

Originally, we attempted to take a snapshot of the GDELT data set and store it in a MySQL database. The combined events, event mentions, and gkg tables consumed over 12TB. These tables proved unwieldy, in terms of query speed. We created table indexes in order to improve query speed, but the data preparation steps described in section 4.2 and the filter operations described in section 4.3 did not satisfy our end-user latency requirements. Additionally, there was a larger cost burden associated with storing this large

data set and maintaining it up-to-date in our own computing infrastructure when compared to using BigQuery in GCP. In the latter setup, data storage costs are free, and the database is automatically updated and maintained by the GDELT organization. There is a cost associated with performing the daily data preparation steps outlined in section 4.2, but these can be anticipated are cheaper than the MySQL solution.

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