Office for National Statistics

Global Database of Events, Language and Tone (GDELT) appendix

Appendix to the main article detailing some technical details for anyone interested in using Global Database of Events, Language and Tone (GDELT) data.

Contact:
Susan Williams
susan.williams@ons.gov.uk
+44 (0)20 7592 8637

Release date:
9 January 2020

Next release:
To be announced

Table of contents

1 . Other pages in this release

This release is split into an article and two accompanying notes.

This appendix provides technical details for anyone interested in using Global Database of Events, Language and Tone (GDELT) data, including an overview of data access options, relevant databases, main variables and examples of inaccuracies discovered in the data that should be considered when using GDELT.

The other pages are:

- a main Article, which presents the findings of the investigation and discusses GDELT data potential, limitations and quality, and outlines suggestions for additional research and different applications
- a Data quality note, which addresses some of the more general quality questions relevant to the use of unofficial data sources in the context of disaster reporting

2 . GDELT databases

The scope of the Global Database of Events, Language and Tone (GDELT) project is constantly evolving making it infeasible to explore all types of information stored.

As of 1 February 2019, GDELT's website lists seven different ways to access different aspects of the data:

- GDELT 2.0 Event Database
- GDELT 2.0 Global Knowledge Graph (GKG)
- GDELT 2.0 Mentions
- GDELT Visual Global Knowledge Graph
- GDELT GKG Special Collections
- GDELT Global Frontpage Graph
- GDELT Summary + GDELT APIs
- GDELT Global Difference Graph

Given the time restrictions of this exploratory project, only the first two collections were explored, with the second being explored in detail.

Versions of GDELT

GDELT 1.0 Event Database

GDELT was founded in 1994 and the 1.0 version of the Event Database covers data from 1979 onwards. Data from 1979 to 2005 is available in the form of one zip file per year. The file size gradually increased from 14.3 MB in 1979 to 125.9 MB in 2005, reflecting the increase in the number of news media, and the frequency and comprehensiveness of event recording.

Data files from January 2006 to March 2013 are available at monthly granularity, with the zipped file size rising from 11 MB in January 2006 to 103.2 MB in March 2013. Data files from 1 April 2013 onwards are available at a daily granularity.

GDELT 1.0 Global Knowledge Graph (GKG)

Coinciding with the event database providing daily data on 1 April 2013, the 1.0 version of the GKG database commenced, providing daily data records.

GDELT 2.0

In February 2015, GDELT version 1.0 was complemented by GDELT 2.0. While version 1.0 was updating once a day, GDELT 2.0 now updates every fifteen minutes, including extensive emotion measures and coverage from 65 live translated languages, as well as a range of other features. While most of the basic features of version 1.0 are continued, new projects and collections use GDELT 2.0.

For this exploration, only GDELT 2.0 has been considered. This meant limiting the time scope to going back no further than February 2015. GDELT 1.0 data can be brought in to reach further back.

Overview of GDELT's Event Database versus Global Knowledge Graph (GKG)

GDELT has two main databases: the Event Database and the Global Knowledge Graph (GKG). The Event Database records individual events, aggregating data from many different news articles.

The GKG records individual news articles, extracting detailed information on every person, location, number and theme mentioned in an article.

GDELT 2.0 Event Database

Entries in the Event Database are made at the level of an individual event following the format: who did what to whom and how many news articles are talking about it. An event captures aspects of a news article, so the same news article can be referenced in several events with different aspects highlighted.

The basic format of an entry always has the following fields (and many more):

Table 1: Basic format of an entry in the Event Database

DATE	ACTOR1	ACTOR2	EVENTCODE	ARTICLECOUNT	TONE	GOLDSTEIN
201902011400 (2pm on February 1)	UK	EU	EXPRESS INTENT TO COOPERATE	120	2.4	3.0

- “ACTOR1” can be individual people, organisations or even countries. It indicates “who” is involved with doing something to “ACTOR2”.

- Events are assigned an “EVENTCODE” following a coding framework, called the Conflict and Mediation Event Observations (CAMEO). It is mainly used in the context of political and social sciences, often to analyse political news and violence. At the highest level, an event falls into one of four categories: material conflict, material cooperation, verbal conflict, or verbal cooperation, splitting into twenty event codes (for example, “Make Public Statement”, “Appeal”, “Demand”, “Reduce Relations”, “Fight” and so on.

- “ARTICLECOUNT” is the total number of documents containing a mention of this event during the 15-minute update in which it was first seen.

- “TONE” is the average sentiment of all documents containing a mention of this event during the 15-minute update in which it was first seen. Tone ranges from negative 100 to positive 100, but most articles are between negative 10 and positive 10.

- “GOLDSTEIN” captures the theoretical potential impact that type of event will have on the stability of a country. This variable depends solely on the “EVENTCODE”. Depending on the intensity of conflict or cooperation inherent in different types of international events, a score between negative 10 and positive 10 is assigned. Regardless of how many people are affected, the same score will be assigned to a protest involving 10 demonstrators or 10,000 demonstrators. As an example “EVENTCODE” = “MILITARY ATTACK; CLASH; ASSAULT” then “GOLDSTEIN” = “-10”.

The Event Database codebook (PDF, 372.69KB) gives an overview of the fields included in the Event Database files and their more detailed descriptions.

Goldstein: this scale depends on the CAMEO Event type and does not consider any further context.

GDELT 2.0 Global Knowledge Graph (GKG)

Entries follow a format of breaking a news article down into details: a news article, all locations, all people, all organisations, all themes, all numbers, all pictures, all sentiments, and all words from specific dictionaries that are used in the article. An overview of all GKG variables and how they are coded can be found in the GKG codebook (PDF, 372.69KB).

Table 2 shows an extract from the GKG for a news article about Storm Desmond that hit the UK during 4 to 6 December 2015, destroying thousands of homes and killing three people. For readability, the extract has been reduced to main variables and the variable content has been cut off after around 40 characters.

Table 2: Example news article entry extracted from the GDELT's Global Knowledge Graph (GKG) database: subset of fields used in our analysis.

V2SOURCECOMMONNAME	V2COUNTS	V1THEMES	V1LOCATIONS	V1PERSONS	V1ORGANIZATIONS
www.leighjournal.co.uk/news	AFFECT#60000000#People#United Kingdom; etc	TAX_FNCACT_MAN; KILL; CRISISLEX_T03_DEAD; etc.	4#Caernarfon, Gwynedd, United Kingdom;	adrian holme; etc.	cumbria police; co antrim; etc.
V1.5TONE	V2GCAM	V2.1ALLNAMES	V2.1AMOUNTS	V1.5TONE	
4.60921, 2.00400, 0.0, etc.	wc:969, c1.3:3, c12.1:56, etc.	Storm Desmond, 35;Environment, 47; etc.	130,flood warnings, 144; etc.	4.60921, 2.00400, 0.0, etc.	

Source: GDELT project

The variables in Table 2 are briefly explained as:

V2SOURCECOMMONNAME: the Uniform Resource Locator (URL) for the top level domain of an article.

V2COUNTS: the list of counts found in the article. Each count entry will have the following form:

- Count Type which is most commonly "Affect", "Arrest", "Kidnap", "Kill", "Protest", "Seize" or "Wound" as specified in CAMEO.
- Count, which is the actual count being reported; in our example, 60 million.
- Object Type, which identifies what the number relates to; in our example, people, although not all GKG counts have an object type associated to them.
- Location, which is GDELT's derivation of the place to which the article mainly refers.

V1THEMES: lists all the themes (XLS, 27KB) that GDELT has mapped the article's text into. Again, these themes are mainly based on CAMEO. Our example only shows a subset of the themes for this article.

There are thousands of different GKG themes, but no complete list seems to be provided anywhere. Probably the most comprehensive list of themes can be found within the documentation section of the GDELT project website. The list currently includes over 2,500 different themes. Among these themes are around 2,200 themes from the World Bank Taxonomy.

V1LOCATIONS: lists all locations mentioned in the article, extracted through the Laetare (2012) algorithm. Entries contain several location attributes, including location type (for example, city, state), full name (city or landmark, state, country), country code (two-character FIPS10-4 country code, for example, "UK"), latitude and longitude.

V1PERSONS: lists all person names found in the text.

V1ORGANIZATIONS: lists all company and organisation names found in the text.

V1.5TONE: a set of six core emotional dimensions found in the text of the article, each delimited by a comma. The main three are:

- Tone, which is an average tone of the document as a whole; the score ranges from negative 100 to positive 100
- Positive Score, representing the percentage of words that were found to be positive
- Negative Score, representing the percentage of words that were found to be negative

GCAM: the Global Content Analysis Measures (GCAM) system is a rich source of information which provides the number of words that are found in a wide collection of data dictionaries. In our example, “wc:969” tells us there are 969 words in the article, “c1.3:3” tells us there are three words in the dictionary coded as c1.3 and “c12.1:56” tells us there are 56 words associated to the dictionary coded as c12.1.

V2.1ALLNAMES: lists all proper names referenced in the document along with the character offsets of where they can be found in the article.

V2.1AMOUNTS: lists all precise numeric amounts found in the article, each with the following form:

- Amount, which is the actual number found
- Object, which identifies the object the number refers to
- Offset, which is the character offset of the quoted statement within the article

Notes for: GDELT databases

Leetaru, K. (2012). Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. D-Lib Magazine, 18(9/10).

3 . GDELT data access details

Data access for the Global Database of Events, Language and Tone (GDELT) 2.0 Event Database and the Global Knowledge Graph (GKG) is possible in a number of ways.

Full text api

- Very easy to access directly through any internet browser’s address bar.
- Can produce different kinds of visualisations almost instantaneously.
- This is only possible for a moving past-three-months window; users who want to go further back have to take another route.

Examples:

Coverage (article volume) of Brexit over the past three months

Map of source countries most often talking about Trump

News images identified by Google's Cloud Vision API as containing imagery of “flood OR rain OR storm”

Google BigQuery

- Allows users to query, export, and conduct modelling of the entire dataset using standard SQL, in near real-time.
- Only very minor queries are possible in the free version’s quota limit; to make proper use of this option, a paid version of BigQuery would be needed.

Raw Data Files

Advanced users and those with unique use cases can download the entire underlying event and graph datasets in CSV format – the data amounts to over 2.5TB 2015 alone.

There are two main options for downloading the raw data files:

1. Using a bespoke R or Python package to combine downloading, importing and pre-processing of the data. This is a great way for exploratory analysis or the analysis of smaller time frames requiring less data. These packages take care of all the pre-processing and load data into the programme in a format that allows immediate analysis.
2. Directly downloading the zipped CSV files and building a dataset or database from them. This is best used when several analyses on the data are planned and a robust downloading approach is preferred. After the downloading, substantial pre-processing is needed before data can be analysed.

For the initial exploration phase, the first approach (downloading through the gdeltr2 R package) was chosen. Beyond initial exploration, direct download was preferred as the omission of pre-processing made the downloading process less resource-intensive and more robust.

4 . Known disaster databases

Three main global databases of multi-disaster loss and damage are:

- EM-DAT
- NatCat-SERVICE
- Sigma

Additionally, the UN database DesInventar is set up to become a global multi-disaster loss and damage database, but it currently still is in development and data coverage is very patchy.

While NatCat-Service and Sigma are maintained by reinsurers and provide limited data access, access to EM-DAT data is free up to a specific threshold. This database is maintained by the Centre for Research on the Epidemiology of Disasters (CRED) of the Université catholique de Louvain.

Comparing these three main global databases is difficult because of inconsistent reporting between them. Since 2009, EM-DATA and NatCat-SERVICE have moved to more consistent reporting standards for natural, but not technological disasters. Even so, a 2018 paper found a statistically significant difference in the descriptive statistics of annual cyclone damages across the Chinese government's databases, EM-DAT and NatCat-SERVICE.

All three databases contain the same overall information such as economic losses, fatalities, numbers injured and affected, damage to infrastructure and buildings. However, the focus of EM-Dat is primarily on the humanitarian aspects, whereas the two reinsurers concentrate more on accurately reecting the material losses (Wirtz et al., 2014). The three databases also apply different documentation thresholds (see event entry criteria in the Overview of main disaster and event databases table).

Main limitations across the databases are a lack of disaggregated data, limited spatial coverage and resolution, insufficient completeness and reliability of data, and insufficient capture of holistic disaster losses, including indirect losses (Moriyama et al., 2018).

Notes for: Known disaster databases

Wirtz, A., Kron, W., Löw, P. et al. The need for data: natural disasters and the challenges of database management. Nat Hazards 70, 135–157 (2014)

Bakkensen, L.A., Shi, X. & Zurita, B.D. EconDisCliCha (2018) 2: 49

Moriyama, K., Sasaki, D., & Ono, Y. (2018). Comparison of Global Databases for Disaster Loss and Damage Data. Journal of Disaster Research, 13(6), 1007-1014