THE GDELT EVENT DATABASE
DATA FORMAT CODEBOOK V2.0
2/19/2015

http://gdeltproject.org/

INTRODUCTION

This codebook provides a quick overview of the fields in the GDELT Event file format and their descriptions. GDELT Event records are stored in an expanded version of the dyadic CAMEO format, capturing two actors and the action performed by Actor1 upon Actor2. A wide array of variables break out the raw CAMEO actor codes into their respective fields to make it easier to interact with the data, the Action codes are broken out into their hierarchy, the Goldstein ranking score is provided, a unique array of georeferencing fields offer estimated landmark-centroid-level geographic positioning of both actors and the location of the action, and a new "Mentions" table records the network trajectory of the story of each event "in flight" through the global media system.

At present, only records from February 19, 2015 onwards are available in the GDELT 2.0 file format, however in late Spring 2015 the entire historical backfile back to 1979 will be released in the GDELT 2.0 format. The Records are stored one per line, separated by a newline (\n) and are tab-delimited (note that files have a ".csv" extension, but are actually tab-delimited).

With the release of GDELT 2.0, the daily GDELT 1.0 Event files will still be generated each morning at least through the end of Spring 2015 to enable existing applications to continue to function without modification. Please note that at present, since GDELT 2.0 files are only available for events beginning February 19, 2015, you will need to use GDELT 1.0 to examine longitudinal patterns (since it stretches back to January 1, 1979) and use GDELT 2.0 moving forward for realtime events.

There are now two data tables created every 15 minutes for the GDELT Event dataset. The first is the traditional Event table. This table is largely identical to the GDELT 1.0 format, but does have several changes as noted below. In addition to the Event table there is now a new Mentions table that records all mentions of each event. As an event is mentioned across multiple news reports, each of those mentions is recorded in the Mentions table, along with several key indicators about that mention, including the location within the article where the mention appeared (in the lead paragraph versus being buried at the bottom) and the "confidence" of the algorithms in their identification of the event from that specific news report. The Confidence measure is a new feature in GDELT 2.0 that makes it possible to adjust the sensitivity of GDELT towards specific use cases. Those wishing to find the earliest glimmers of breaking events or reports of very small-bore events that tend to only appear as part of period "round up" reports, can use the entire event stream, while those wishing to find only the largest events with strongly detailed descriptions, can filter the Event stream to find only those events with the highest Confidence measures. This allows the GDELT Event stream to be dynamically filtered for each individual use case (learn more about the Confidence measure below). It also makes it possible to identify the "best" news report to return for a given event (filtering all mentions of an event for those with the highest Confidence scores, most prominent positioning within the article, and/or in a specific source language – such as Arabic coverage of a protest versus English coverage of that protest).

EVENT TABLE

EVENTID AND DATE ATTRIBUTES

The first few fields of an event record capture its globally unique identifier number, the date the event took place on, and several alternatively formatted versions of the date designed to make it easier to work with the event records in different analytical software programs that may have specific date format requirements. The parenthetical after each variable name gives the datatype of that field.

Note that even though GDELT 2.0 operates at a 15 minute resolution, the date fields in this section still record the date at the daily level, since this is the resolution that event analysis has historically been performed at. To examine events at the 15 minute resolution, use the DATEADDED field (the second from the last field in this table at the end).

- **GlobalEventID.** (integer) Globally unique identifier assigned to each event record that uniquely identifies it in the master dataset. **NOTE**: While these will often be sequential with date, this is NOT always the case and this field should NOT be used to sort events by date: the date fields should be used for this. **NOTE:** There is a large gap in the sequence between February 18, 2015 and February 19, 2015 with the switchover to GDELT 2.0 – these are not missing events, the ID sequence was simply reset at a higher number so that it is possible to easily distinguish events created after the switchover to GDELT 2.0 from those created using the older GDELT 1.0 system.
- **Day**. (integer) Date the event took place in YYYYMMDD format. See DATEADDED field for YYYYMMDDHHMMSS date.
- **MonthYear.** (integer) Alternative formatting of the event date, in YYYYMM format.
- **Year.** (integer) Alternative formatting of the event date, in YYYY format.
- **FractionDate.** (floating point) Alternative formatting of the event date, computed as YYYY.FFFF, where FFFF is the percentage of the year completed by that day. This collapses the month and day into a fractional range from 0 to 0.9999, capturing the 365 days of the year. The fractional component (FFFF) is computed as (MONTH * 30 + DAY) / 365. This is an approximation and does not correctly take into account the differing numbers of days in each month or leap years, but offers a simple single-number sorting mechanism for applications that wish to estimate the rough temporal distance between dates.

ACTOR ATTRIBUTES

The next fields describe attributes and characteristics of the two actors involved in the event. This includes the complete raw CAMEO code for each actor, its proper name, and associated attributes. The raw CAMEO code for each actor contains an array of coded attributes indicating geographic, ethnic, and religious affiliation and the actor's role in the environment (political elite, military officer, rebel, etc). These 3-character codes may be combined in any order and are concatenated together to form the final raw actor CAMEO code. To make it easier to utilize this information in analysis, this section breaks these codes out into a set of individual fields that can be separately queried. **NOTE**: all attributes in this section other than CountryCode are derived from the TABARI ACTORS dictionary and are NOT supplemented from information in the text. Thus, if the text refers to a group as "Radicalized terrorists," but the TABARI ACTORS dictionary labels that group as "Insurgents," the latter label will be used. Use the GDELT Global Knowledge Graph to enrich actors with additional information from the rest of the article. **NOTE:** the CountryCode field reflects a combination of information from the TABARI ACTORS dictionary and text, with the ACTORS dictionary taking precedence, and thus if the text refers to "French Assistant Minister Smith was in Moscow," the CountryCode field will list France in the CountryCode field, while the geographic fields discussed at the end of this manual may list Moscow as his/her location. **NOTE**: One of the two actor fields may be blank in complex or single-actor situations or may contain only minimal detail for actors such as "Unidentified gunmen."

GDELT currently uses the CAMEO version 1.1b3 taxonomy. For more information on what each specific code in the fields below stands for and the complete available taxonomy of the various fields below, please see the CAMEO User Manual [1] or the GDELT website for crosswalk files.[2]

- **Actor1Code**. (string) The complete raw CAMEO code for Actor1 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor1.
- **Actor1Name**. (string) The actual name of the Actor1. In the case of a political leader or organization, this will be the leader's formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor1.
- **Actor1CountryCode**. (string) The 3-character CAMEO code for the country affiliation of Actor1. May be blank if the system was unable to identify an Actor1 or determine its country affiliation (such as "UNIDENTIFIED GUNMEN").
- **Actor1KnownGroupCode.** (string) If Actor1 is a known IGO/NGO/rebel organization (United Nations, World Bank, al-Qaeda, etc) with its own CAMEO code, this field will contain that code.
- **Actor1EthnicCode**. (string) If the source document specifies the ethnic affiliation of Actor1 and that ethnic group has a CAMEO entry, the CAMEO code is entered here. **NOTE**: a few special groups like ARAB may also have entries in the type column due to legacy CAMEO behavior. **NOTE:** this behavior is highly experimental and may not capture all affiliations properly – for more comprehensive and sophisticated identification of ethnic affiliation, it is recommended that users use the GDELT Global Knowledge Graph's ethnic, religious, and social group taxonomies and post-enrich actors from the GKG.
- **Actor1Religion1Code**. (string) If the source document specifies the religious affiliation of Actor1 and that religious group has a CAMEO entry, the CAMEO code is entered here. **NOTE**: a few special groups like JEW may also have entries in the geographic or type columns due to legacy CAMEO behavior. **NOTE:** this behavior is highly experimental and may not capture all affiliations properly – for more comprehensive and sophisticated identification of ethnic affiliation, it is recommended that users use the GDELT Global Knowledge Graph's ethnic, religious, and social group taxonomies and post-enrich actors from the GKG.
- **Actor1Religion2Code.** (string) If multiple religious codes are specified for Actor1, this contains the secondary code. Some religion entries automatically use two codes, such as Catholic, which invokes Christianity as Code1 and Catholicism as Code2.
- **Actor1Type1Code**. (string) The 3-character CAMEO code of the CAMEO "type" or "role" of Actor1, if specified. This can be a specific role such as Police Forces, Government, Military, Political Opposition, Rebels, etc, a broad role class such as Education, Elites, Media, Refugees, or

[1] http://gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf
[2] http://gdeltproject.org/

organizational classes like Non-Governmental Movement. Special codes such as Moderate and Radical may refer to the operational strategy of a group.

- **Actor1Type2Code.** (string) If multiple type/role codes are specified for Actor1, this returns the second code.
- **Actor1Type3Code.** (string) If multiple type/role codes are specified for Actor1, this returns the third code.

The fields above are repeated for **Actor2**. The set of fields above are repeated, but each is prefaced with "Actor2" instead of "Actor1". The definitions and values of each field are the same as above.

EVENT ACTION ATTRIBUTES

The following fields break out various attributes of the event "action" (what Actor1 did to Actor2) and offer several mechanisms for assessing the "importance" or immediate-term "impact" of an event. **NOTE**: the various fields in this section recording the amount of coverage an event has received are included solely for legacy purposes – the new Mentions table should be used instead in most cases.

- **IsRootEvent**. (integer) The system codes every event found in an entire document, using an array of techniques to deference and link information together. A number of previous projects such as the ICEWS initiative have found that events occurring in the lead paragraph of a document tend to be the most "important." This flag can therefore be used as a proxy for the rough importance of an event to create subsets of the event stream. **NOTE**: this field refers only to the first news report to mention an event and is not updated if the event is found in a different context in other news reports. It is included for legacy purposes – for more precise information on the positioning of an event, see the Mentions table.
- **EventCode**. (string) This is the raw CAMEO action code describing the action that Actor1 performed upon Actor2. **NOTE**: it is strongly recommended that this field be stored as a string instead of an integer, since the CAMEO taxonomy can include zero-leaded event codes that can make distinguishing between certain event types more difficult when stored as an integer.
- **EventBaseCode**. (string) CAMEO event codes are defined in a three-level taxonomy. For events at level three in the taxonomy, this yields its level two leaf root node. For example, code "0251" ("Appeal for easing of administrative sanctions") would yield an EventBaseCode of "025" ("Appeal to yield"). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode. **NOTE**: it is strongly recommended that this field be stored as a string instead of an integer, since the CAMEO taxonomy can include zero-leaded event codes that can make distinguishing between certain event types more difficult when stored as an integer.
- **EventRootCode**. (string) Similar to EventBaseCode, this defines the root-level category the event code falls under. For example, code "0251" ("Appeal for easing of administrative sanctions") has a root code of "02" ("Appeal"). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode. **NOTE**: it is strongly recommended that this field be stored as a string instead of an integer, since the CAMEO taxonomy can include zero-leaded event codes that can make distinguishing between certain event types more difficult when stored as an integer.
- **QuadClass.** (integer) The entire CAMEO event taxonomy is ultimately organized under four primary classifications: Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material

Conflict. This field specifies this primary classification for the event type, allowing analysis at the highest level of aggregation. The numeric codes in this field map to the Quad Classes as follows: 1=Verbal Cooperation, 2=Material Cooperation, 3=Verbal Conflict, 4=Material Conflict.

- **GoldsteinScale.** (floating point) Each CAMEO event code is assigned a numeric score from -10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country. This is known as the Goldstein Scale. This field specifies the Goldstein score for each event type. **NOTE:** this score is based on the type of event, not the specifics of the actual event record being recorded – thus two riots, one with 10 people and one with 10,000, will both receive the same Goldstein score. This can be aggregated to various levels of time resolution to yield an approximation of the stability of a location over time.
- **NumMentions.** (integer) This is the total number of mentions of this event across all source documents **during the 15 minute update in which it was first seen**. Multiple references to an event within a single document also contribute to this count. This can be used as a method of assessing the "importance" of an event: the more discussion of that event, the more likely it is to be significant. The total universe of source documents and the density of events within them vary over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest. This field is actually a composite score of the total number of raw mentions and the number of mentions extracted from reprocessed versions of each article (see the discussion for the Mentions table). **NOTE**: this field refers only to the first news report to mention an event and is not updated if the event is found in a different context in other news reports. It is included for legacy purposes – for more precise information on the positioning of an event, see the Mentions table.
- **NumSources**. (integer) This is the total number of information sources containing one or more mentions of this event **during the 15 minute update in which it was first seen**. This can be used as a method of assessing the "importance" of an event: the more discussion of that event, the more likely it is to be significant. The total universe of sources varies over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest. **NOTE**: this field refers only to the first news report to mention an event and is not updated if the event is found in a different context in other news reports. It is included for legacy purposes – for more precise information on the positioning of an event, see the Mentions table.
- **NumArticles.** (integer) This is the total number of source documents containing one or more mentions of this event **during the 15 minute update in which it was first seen**. This can be used as a method of assessing the "importance" of an event: the more discussion of that event, the more likely it is to be significant. The total universe of source documents varies over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest. **NOTE**: this field refers only to the first news report to mention an event and is not updated if the event is found in a different context in other news reports. It is included for legacy purposes – for more precise information on the positioning of an event, see the Mentions table.
- **AvgTone.** (numeric) This is the average "tone" of all documents containing one or more mentions of this event **during the 15 minute update in which it was first seen**. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This can be used as a method of filtering the "context" of events as a subtle measure of the importance of an event and as a proxy for the "impact" of that event. For example, a riot event with a slightly negative average tone is likely to have been a minor occurrence, whereas if it had an extremely negative average tone, it suggests a far more serious occurrence. A riot with a positive score likely suggests a very minor

occurrence described in the context of a more positive narrative (such as a report of an attack occurring in a discussion of improving conditions on the ground in a country and how the number of attacks per day has been greatly reduced). **NOTE**: this field refers only to the first news report to mention an event and is not updated if the event is found in a different context in other news reports. It is included for legacy purposes – for more precise information on the positioning of an event, see the Mentions table. **NOTE:** this provides only a basic tonal assessment of an article and it is recommended that users interested in emotional measures use the Mentions and Global Knowledge Graph tables to merge the complete set of 2,300 emotions and themes from the GKG GCAM system into their analysis of event records.

EVENT GEOGRAPHY

The final set of fields add a novel enhancement to the CAMEO taxonomy, georeferencing each event along three primary dimensions to the landmark-centroid level. To do this, the fulltext of the source document is processed using fulltext geocoding and automatic disambiguation to identify every geographic reference.[3] The closest reference to each of the two actors and to the action reference are then encoded in these fields. The georeferenced location for an actor may not always match the Actor1_CountryCode or Actor2_CountryCode field, such as in a case where the President of Russia is visiting Washington, DC in the United States, in which case the Actor1_CountryCode would contain the code for Russia, while the georeferencing fields below would contain a match for Washington, DC. It may not always be possible for the system to locate a match for each actor or location, in which case one or more of the fields may be blank. The Action fields capture the location information closest to the point in the event description that contains the actual statement of action and is the best location to use for placing events on a map or in other spatial context.

To find all events located in or relating to a specific city or geographic landmark, the Geo_FeatureID column should be used, rather than the Geo_Fullname column. This is because the Geo_Fullname column captures the name of the location as expressed in the text and thus reflects differences in transliteration, alternative spellings, and alternative names for the same location. For example, Mecca is often spelled Makkah, while Jeddah is commonly spelled Jiddah or Jaddah. The Geo_Fullname column will reflect each of these different spellings, while the Geo_FeatureID column will resolve them all to the same unique GNS or GNIS feature identification number. For more information on the GNS and GNIS identifiers, see Leetaru (2012). [4]

When looking for events in or relating to a specific country, such as Syria, there are two possible filtering methods. The first is to use the Actor_CountryCode fields in the Actor section to look for all actors having the SYR (Syria) code. However, conflict zones are often accompanied by high degrees of uncertainty in media reporting and a news article might mention only “Unidentified gunmen stormed a house and shot 12 civilians.” In this case, the Actor_CountryCode fields for Actor1 and Actor2 would both be blank, since the article did not specify the actor country affiliations, while their Geo_CountryCode values (and the ActorGeo_CountryCode for the event) would specify Syria. This can result in dramatic differences when examining active conflict zones. The second method is to examine the ActorGeo_CountryCode for the location of the event. This will also capture situations such as the United States criticizing a statement by Russia regarding a specific Syrian attack.

[3] http://www.dlib.org/dlib/september12/leetaru/09leetaru.html
[4] http://www.dlib.org/dlib/september12/leetaru/09leetaru.html

- **Actor1Geo_Type**. (integer) This field specifies the geographic resolution of the match type and holds one of the following values: 1=COUNTRY (match was at the country level), 2=USSTATE (match was to a US state), 3=USCITY (match was to a US city or landmark), 4=WORLDCITY (match was to a city or landmark outside the US), 5=WORLDSTATE (match was to an Administrative Division 1 outside the US – roughly equivalent to a US state). This can be used to filter events by geographic specificity, for example, extracting only those events with a landmark-level geographic resolution for mapping. Note that matches with codes 1 (COUNTRY), 2 (USSTATE), and 5 (WORLDSTATE) will still provide a latitude/longitude pair, which will be the centroid of that country or state, but the FeatureID field below will be blank.
- **Actor1Geo_Fullname**. (string) This is the full human-readable name of the matched location. In the case of a country it is simply the country name. For US and World states it is in the format of "State, Country Name", while for all other matches it is in the format of "City/Landmark, State, Country". This can be used to label locations when placing events on a map. **NOTE**: this field reflects the precise name used to refer to the location in the text itself, meaning it may contain multiple spellings of the same location – use the FeatureID column to determine whether two location names refer to the same place.
- **Actor1Geo_CountryCode**. (string) This is the 2-character FIPS10-4 country code for the location.
- **Actor1Geo_ADM1Code**. (string). This is the 2-character FIPS10-4 country code followed by the 2-character FIPS10-4 administrative division 1 (ADM1) code for the administrative division housing the landmark. In the case of the United States, this is the 2-character shortform of the state's name (such as "TX" for Texas).
- **Actor1Geo_ADM2Code**. (string). For international locations this is the numeric Global Administrative Unit Layers (GAUL) administrative division 2 (ADM2) code assigned to each global location, while for US locations this is the two-character shortform of the state's name (such as "TX" for Texas) followed by the 3-digit numeric county code (following the INCITS 31:200x standard used in GNIS). For more detail on the contents and computation of this field, please see the following footnoted URL. [5] **NOTE**: This field may be blank/null in cases where no ADM2 information was available, for some ADM1-level matches, and for all country-level matches. **NOTE:** this field may still contain a value for ADM1-level matches depending on how they are codified in GNS.
- **Actor1Geo_Lat**. (floating point) This is the centroid latitude of the landmark for mapping.
- **Actor1Geo_Long**. (floating point) This is the centroid longitude of the landmark for mapping.
- **Actor1Geo_FeatureID**. (string). This is the GNS or GNIS FeatureID for this location. More information on these values can be found in Leetaru (2012).[6] **NOTE:** When Actor1Geo_Type has a value of 3 or 4 this field will contain a signed numeric value, while it will contain a textual FeatureID in the case of other match resolutions (usually the country code or country code and ADM1 code). A small percentage of small cities and towns may have a blank value in this field even for Actor1Geo_Type values of 3 or 4: this will be corrected in the 2.0 release of GDELT. **NOTE**: This field can contain both positive and negative numbers, see Leetaru (2012) for more information on this.

These codes are repeated for **Actor2** and **Action**, using those prefixes.

DATA MANAGEMENT FIELDS

[5] http://blog.gdeltproject.org/global-second-order-administrative-divisions-now-available-from-gaul/
[6] http://www.dlib.org/dlib/september12/leetaru/09leetaru.html

Finally, a set of fields at the end of the record provide additional data management information for the event record.

- **DATEADDED**. (integer) This field stores the date the event was added to the master database in YYYYMMDDHHMMSS format in the UTC timezone. For those needing to access events at 15 minute resolution, this is the field that should be used in queries.
- **SOURCEURL**. (string) This field records the URL or citation of the first news report it found this event in. In most cases this is the first report it saw the article in, but due to the timing and flow of news reports through the processing pipeline, this may not always be the very first report, but is at least in the first few reports.

MENTIONS TABLE

The Mentions table is a new addition to GDELT 2.0 and records each mention of the events in the Event table, making it possible to track the trajectory and network structure of a story as it flows through the global media system. Each mention of an event receives its own entry in the Mentions table – thus an event which is mentioned in 100 articles will be listed 100 times in the Mentions table. Mentions are recorded irrespective of the date of the original event, meaning that a mention today of an event from a year ago will still be recorded, making it possible to trace discussion of "anniversary events" or historical events being recontextualized into present actions. If a news report mentions multiple events, each mention is recorded separately in this table. For translated documents, all measures below are based on its English translation.

Several of the new measures recorded in the Mentions table make it possible to better filter events based on how confident GDELT was in its extraction of that event. When trying to understand news media spanning the entire globe, one finds that journalism is rife with ambiguities, assumed background knowledge, and complex linguistic structures. Not every event mention will take the form of "American President Barack Obama met with Russian President Vladimir Putin yesterday at a trade summit in Paris, France." Instead, an event mention might more commonly appear as "Obama and Putin were in Paris yesterday for a trade summit. The two leaders met backstage where he discussed his policy on Ukraine." To which of the two leader(s) do "he" and "his" refer? Is Obama discussing Obama's policy on Ukraine, or is Obama discussing Putin's policy on Ukraine, or is it Putin discussing Putin's policy or perhaps Putin discussing Obama's policy? While additional cues may be available in the surrounding text, ambiguous event mentions like this are exceptionally common across the world's media. Similarly, it would be difficult indeed to maintain an exhaustive list of every single political figure in the entire world and thus context is often critical for disambiguating the geographic affiliation of an actor. Even in the case of more senior political leadership, a reference to "Renauld's press conference this afternoon in Port-au-Prince" most likely refers to Lener Renauld, the Minister of Defense of Haiti, but this disambiguation still carries with it some degree of ambiguity. GDELT makes use of an array of natural language processing algorithms like coreference and deep parsing using whole-of-document context. While these enormously increase GDELT's ability to understand and extract ambiguous and linguistically complex events, such extractions also come with a higher potential for error. Under GDELT 1.0, the NumMentions field as designed as a composite score of the absolute number of unique documents mentioning an event and the number of revisions to the text required by these various algorithms, up to six revision passes. Under GDELT 2.0, the Mentions table now separates these, with each record in the Mentions table recording an individual mention of an event in an article, while the new Confidence field records GDELT’s confidence in its extraction of that event from that particular article. This field is a percent, ranging from 10 to 100% and indicates how aggressively GDELT had to perform tasks like coreference or grammatical restructuring to extract the event from that article. Sorting all mentions of an event by this field makes it possible to identify articles featuring the strongest and most unambiguous discussion of an event.

- **GlobalEventID.** (integer) This is the ID of the event that was mentioned in the article.
- **EventTimeDate.** (integer) This is the 15-minute timestamp (YYYYMMDDHHMMSS) when the event being mentioned was first recorded by GDELT (the DATEADDED field of the original event record). This field can be compared against the next one to identify events being mentioned for the first time (their first mentions) or to identify events of a particular vintage being mentioned now (such as filtering for mentions of events at least one week old).
- **MentionTimeDate.** (integer) This is the 15-minute timestamp (YYYYMMDDHHMMSS) of the current update. This is identical for all entries in the update file but is included to make it easier to load the Mentions table into a database.
- **MentionType.** (integer) This is a numeric identifier that refers to the source collection the document came from and is used to interpret the MentionIdentifier in the next column. In essence, it specifies how to interpret the MentionIdentifier to locate the actual document. At present, it can hold one of the following values:
  - 1 = WEB (The document originates from the open web and the MentionIdentifier is a fully-qualified URL that can be used to access the document on the web).
  - 2 = CITATIONONLY (The document originates from a broadcast, print, or other offline source in which only a textual citation is available for the document. In this case the MentionIdentifier contains the textual citation for the document).
  - 3 = CORE (The document originates from the CORE archive and the MentionIdentifier contains its DOI, suitable for accessing the original document through the CORE website).
  - 4 = DTIC (The document originates from the DTIC archive and the MentionIdentifier contains its DOI, suitable for accessing the original document through the DTIC website).
  - 5 = JSTOR (The document originates from the JSTOR archive and the MentionIdentifier contains its DOI, suitable for accessing the original document through your JSTOR subscription if your institution subscribes to it).
  - 6 = NONTEXTUALSOURCE (The document originates from a textual proxy (such as closed captioning) of a non-textual information source (such as a video) available via a URL and the MentionIdentifier provides the URL of the non-textual original source. At present, this Collection Identifier is used for processing of the closed captioning streams of the Internet Archive Television News Archive in which each broadcast is available via a URL, but the URL offers access only to the video of the broadcast and does not provide any access to the textual closed captioning used to generate the metadata. This code is used in order to draw a distinction between URL-based textual material (Collection Identifier 1 (WEB) and URL-based non-textual material like the Television News Archive).
- **MentionSourceName**. (integer) This is a human-friendly identifier of the source of the document. For material originating from the open web with a URL this field will contain the top-level domain the page was from. For BBC Monitoring material it will contain “BBC Monitoring” and for JSTOR material it will contain “JSTOR.” This field is intended for human display of major sources as well as for network analysis of information flows by source, obviating the requirement to perform domain or other parsing of the MentionIdentifier field.

- **MentionIdentifier**. (integer) This is the unique external identifier for the source document. It can be used to uniquely identify the document and access it if you have the necessary subscriptions or authorizations and/or the document is public access. This field can contain a range of values, from URLs of open web resources to textual citations of print or broadcast material to DOI identifiers for various document repositories. For example, if MentionType is equal to 1, this field will contain a fully-qualified URL suitable for direct access. If MentionType is equal to 2, this field will contain a textual citation akin to what would appear in an academic journal article referencing that document (NOTE that the actual citation format will vary (usually between APA, Chicago, Harvard, or MLA) depending on a number of factors and no assumptions should be made on its precise format at this time due to the way in which this data is currently provided to GDELT – future efforts will focus on normalization of this field to a standard citation format). If MentionType is 3, the field will contain a numeric or alpha-numeric DOI that can be typed into JSTOR's search engine to access the document if your institution has a JSTOR subscription.
- **SentenceID**. (integer) The sentence within the article where the event was mentioned (starting with the first sentence as 1, the second sentence as 2, the third sentence as 3, and so on). This can be used similarly to the CharOffset fields below, but reports the event's location in the article in terms of sentences instead of characters, which is more amenable to certain measures of the "importance" of an event's positioning within an article.
- **Actor1CharOffset**. (integer) The location within the article (in terms of English characters) where Actor1 was found. This can be used in combination with the GKG or other analysis to identify further characteristics and attributes of the actor. **NOTE:** due to processing performed on each article, this may be slightly offset from the position seen when the article is rendered in a web browser.
- **Actor2CharOffset**. (integer) The location within the article (in terms of English characters) where Actor2 was found. This can be used in combination with the GKG or other analysis to identify further characteristics and attributes of the actor. **NOTE:** due to processing performed on each article, this may be slightly offset from the position seen when the article is rendered in a web browser.
- **ActionCharOffset**. (integer) The location within the article (in terms of English characters) where the core Action description was found. This can be used in combination with the GKG or other analysis to identify further characteristics and attributes of the actor. **NOTE:** due to processing performed on each article, this may be slightly offset from the position seen when the article is rendered in a web browser.
- **InRawText**. (integer) This records whether the event was found in the original unaltered raw article text (a value of 1) or whether advanced natural language processing algorithms were required to synthesize and rewrite the article text to identify the event (a value of 0). See the discussion on the Confidence field below for more details. Mentions with a value of "1" in this field likely represent strong detail-rich references to an event.
- **Confidence**. (integer) Percent confidence in the extraction of this event from this article. See the discussion above.
- **MentionDocLen**. (integer) The length in English characters of the source document (making it possible to filter for short articles focusing on a particular event versus long summary articles that casually mention an event in passing).
- **MentionDocTone**. (integer) The same contents as the AvgTone field in the Events table, but computed for this particular article. **NOTE**: users interested in emotional measures should use

the MentionIdentifier field above to merge the Mentions table with the GKG table to access the complete set of 2,300 emotions and themes from the GCAM system.

- **MentionDocTranslationInfo**. (string) This field is internally delimited by semicolons and is used to record provenance information for machine translated documents indicating the original source language and the citation of the translation system used to translate the document for processing. It will be blank for documents originally in English. At this time the field will also be blank for documents translated by a human translator and provided to GDELT in English (such as BBC Monitoring materials) – in future this field may be expanded to include information on human translation pipelines, but at present it only captures information on machine translated materials. An example of the contents of this field might be "srclc:fra; eng:Moses 2.1.1 / MosesCore Europarl fr-en / GT-FRA 1.0". **NOTE**: Machine translation is often not as accurate as human translation and users requiring the highest possible confidence levels may wish to exclude events whose only mentions are in translated reports, while those needing the highest-possible coverage of the non-Western world will find that these events often offer the earliest glimmers of breaking events or smaller-bore events of less interest to Western media.
    - **SRCLC**. This is the Source Language Code, representing the three-letter ISO639-2 code of the language of the original source material.
    - **ENG**. This is a textual citation string that indicates the engine(s) and model(s) used to translate the text. The format of this field will vary across engines and over time and no expectations should be made on the ordering or formatting of this field. In the example above, the string "Moses 2.1.1 / MosesCore Europarl fr-en / GT-FRA 1.0" indicates that the document was translated using version 2.1.1 of the Moses [7] SMT platform, using the "MosesCore Europarl fr-en" translation and language models, with the final translation enhanced via GDELT Translingual's own version 1.0 French translation and language models. A value of "GT-ARA 1.0" indicates that GDELT Translingual's version 1.0 Arabic translation and language models were the sole resources used for translation. Additional language systems used in the translation pipeline such as word segmentation systems are also captured in this field such that a value of "GT-ZHO 1.0 / Stanford PKU" indicates that the Stanford Chinese Word Segmenter [8] was used to segment the text into individual words and sentences, which were then translated by GDELT Translingual's own version 1.0 Chinese (Traditional or Simplified) translation and language models.
- **Extras**. (string) This field is currently blank, but is reserved for future use to encode special additional measurements for selected material.

[7] http://www.statmt.org/moses/

[8] http://nlp.stanford.edu/software/segmenter.shtml