GDELT – DATA FORMAT CODEBOOK V 1.03
8/25/2013
http://gdeltproject.org/

INTRODUCTION

This codebook provides a quick overview of the fields in the GDELT data 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, an average "tone" score is provided for all coverage of the event, several indicators of "importance" based on media attention are provided, and a unique array of georeferencing fields offer estimated landmark-centroid-level geographic positioning of both actors and the location of the action.

The Historical Backfile collection, which runs January 1, 1979 through March 31, 2013 contains 57 fields for each record. The Daily Updates collection, which begins April 1, 2013 and runs through present, contains an additional field at the end of each record, for a total of 58 fields for each record. 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).

For more information and to download both the historical backfiles and daily event files, please see the main GDELT website at http://gdeltproject.org/

DATA FIELDS

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 formatting (based on R's datatypes) for the variable to facilitate formatting in other software.

- **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 appear to be a very small number of records that received duplicate GLOBALEVENTID values due to an unknown error in the numbering subsystem early on. This did not affect the deduplication process, since that operates directly on the attribute fields themselves, but does appear to have led to some duplicate GlobalEventIDs. These appear to be randomly distributed and the duplicate events should be safe to discard or renumber.
- **Day**. (integer) Date the event took place in YYYYMMDD format.
- **MonthYear.** (integer) Alternative formatting of the event date, in YYYYMM format.
- **Year.** (integer) Alternative formatting of the event date, in YYYY format.
- **FractionDate.** (numeric) 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. **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 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**. (character or factor) 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**. (character) The actual name of the Actor 1. 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**. (character or factor) 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"). Note that through 8/26/2013 matches for South Sudan ("SSD") may be missing from this field. The country code will be correctly present in the raw CAMEO code in Actor1Code, but may not have been correctly placed into the Actor1CountryCode field – only South Sudan was affected by this issue.

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

- **Actor1KnownGroupCode.** (character or factor) 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**. (character or factor) 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 brand-new and highly experimental and may not capture all affiliations properly.
- **Actor1Religion1Code**. (character or factor) 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 brand-new and highly experimental and may not capture all affiliations properly.
- **Actor1Religion2Code. (**character or factor) 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**. (character or factor) 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 organizational classes like Non-Governmental Movement. Special codes such as Moderate and Radical may refer to the operational strategy of a group.
- **Actor1Type2Code.** (character or factor) If multiple type/role codes are specified for Actor1, this returns the second code.
- **Actor1Type3Code.** (character or factor) 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.

- **IsRootEvent**. (logical or binary or byte) 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.
- **EventCode**. (character or factor) This is the raw CAMEO action code describing the action that Actor1 performed upon Actor2.
- **EventBaseCode**. (character or factor) 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.

- **EventRootCode**. (character or factor) 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.
- **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.** (numeric) 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. 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. **NOTE**: this field is updated over time if news articles published later discuss this event (for example, in the weeks after a major bombing there will likely be numerous news articles published mentioning the original bombing as context to new developments, while on the one-year anniversary there will likely be further coverage). At this time the daily event stream only includes new event records found each day and does not include these updates; a special "updates" stream will be released in Fall 2013 that will include these.
- **NumSources**. (integer) This is the total number of information sources containing one or more mentions of this event. 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**: same as with NumMentions, this field is updated over time to reflect subsequent coverage of the event. Similarly, these updates are not included in the daily event stream, but will be incorporated into a new "updates" stream to be released in Fall 2013.
- **NumArticles.** (integer) This is the total number of source documents containing one or more mentions of this event. 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**: same as with NumMentions, this field is updated over time to reflect subsequent coverage of the event, but these updates are not currently part of the daily event stream.

- **AvgTone.** (numeric) This is the average "tone" of all documents containing one or more mentions of this event. 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).

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**. (character) 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**. (character) This is the 2-character FIPS10-4 country code for the location.
- **Actor1Geo_ADM1Code**. (character) 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_Lat**. (numeric) This is the centroid latitude of the landmark for mapping.
- **Actor1Geo_Long**. (numeric) This is the centroid longitude of the landmark for mapping.
- **Actor1Geo_FeatureID**. (signed integer) This is the GNS or GNIS FeatureID for this location. More information on these values can be found in Leetaru (2012).[5] **NOTE:** This field will be blank except when Actor1Geo_Type has a value of 3 or 4. 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

Finally, a set of fields at the end of the record provide data management information for the event record. For events prior to April 1, 2013, the only field in this section is DATEADDED, which gives the date the event was added to the database. This is useful when performing updates against mirrors of the database, since news coverage published today could add events from the distant past, which would result in the SQLDATE and other event date fields containing the date the event actually took place, while the DATEADDED field below will carry today's date. For events on or after April 1, 2013, an additional column, SOURCEURL, lists the URL of the news article the event was found in, or "BBC Monitoring" for articles from the BBC Monitoring service.

- **DATEADDED**. (integer) This field stores the date the event was added to the master database.

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

- **SOURCEURL**. (character) This field is only present in the daily event stream files beginning April 1, 2013 and lists the URL of the news article the event was found in. If the event was found in an article from the BBC Monitoring service, this field will contain "BBC Monitoring." If an event was mentioned in multiple articles, only one of the URLs is provided. This field is not present in event files prior to April 1, 2013.