THE GDELT GLOBAL KNOWLEDGE GRAPH (GKG)
DATA FORMAT CODEBOOK V 1.0ALPHAEXP
(Alpha Release - Experimental)
11/3/2013
http://gdeltproject.org/

INTRODUCTION

This codebook introduces the new GDELT Global Knowledge Graph (GKG), which expands GDELT's ability to quantify global human society beyond cataloging physical occurrences towards actually representing all of the latent dimensions, geography, and network structure of the global news. To sum up the GKG in a single sentence, it connects every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day.

There are actually two separate output streams produced under the auspices of the GKG. The first is the daily Counts File, which records mentions of counts of things with respect to a set of predefined categories such as a number of protesters, a number killed, or a number displaced or sickened. Such counts may occur independently of the CAMEO events in the primary GDELT event stream, such as mentions of those killed in industrial accidents (which are not captured in CAMEO) or those displaced by a natural disaster or sickened by a disease epidemic. In this way, the GKG Counts File can be used to produce a daily "Death Tracker" to map all mentions of death across the world each day, or an "Affected Tracker" to indicate how many persons were sickened/displaced/stranded each day (at least as recorded in the global news media). The second file is the GKG Graph File, which contains the actual graph connecting all persons, organizations, locations, emotions, themes, counts, events, and sources together each day into a single network structure and captures the cultural narratives that envelope the global information stream.

Due to the vast number of use cases articulated for the GKG already, a decision was made to create a raw output format that could be processed into the necessary refined formats for a wide array of software packages and analysis needs and that would support a diverse assortment of extremely complex analytic needs in a single file. Thus, unlike the primary GDELT event stream, which is designed for direct import into major statistical packages like R, the GKG file format requires more sophisticated preprocessing and users will likely want to make use of a scripting language like PERL or Python to extract and reprocess the data for import into a statistical package. Thus, users may require more advanced text processing and scripting language skills to work with the GKG data and additional nuance may be required when thinking about how to incorporate these indicators into statistical models and network and geographic constructs, as outlined in this codebook.

SPECIAL NOTICE – ALPHA EXPERIMENTAL STATUS

Unlike the primary GDELT event stream, the GDELT Global Knowledge Graph is a highly experimental new capability that is still undergoing active development and is currently made available as an **ALPHA EXPERIMENTAL** version release, meaning specifics, especially the output format, may change in the future with the next version release as the system evolves to incorporate community feedback and needs.

DATA FIELDS – COUNT FILES

This section outlines the specific data fields in the Count Files (*.gkgcounts.csv). Unlike the primary GDELT event stream, these records are not issued unique identifier numbers, nor are they dated beyond the date of the news content used to generate the file. Users should note that all Counts are dated with the date of the news media used to identify them, not the actual date of the Count, though future releases of the GKG will emphasize adding date information into the file to handle mentions of counts from the past. As an example of how to interpret this file, an entry with NumArts=10, CountType=KILL, Number=47, ObjectType="jihadists" means that "47 Jihadists were killed" was found in ten different news articles that day.

PRIMARY ATTRIBUTES

These fields capture the primary attributes of the count.

- **DATE**. (integer) This is the date in YYYYMMDD format on which the news media used to construct this GKG file was published. NOTE that unlike the main GDELT event stream files, this date is the date of publication of the news media from which the information was extracted – if the article discusses events in the past, the date is NOT time-shifted as it is for the GDELT event stream. This date will be the same for all rows in a file and is redundant from a data processing standpoint, but is provided to make it easier to load GKG files directly into an SQL database for analysis.
- **NUMARTS**. (integer) This is the total number of source documents containing one or more mentions of this count. This can be used as a method of assessing the "importance" of an count: the more discussion of that count, 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 counts during the time period of interest.
- **COUNTTYPE**. (character) This is the value of the NAME field from the Category List spreadsheet indicating which category this count is of. At the time of this writing, this is most often AFFECT, ARREST, KIDNAP, KILL, PROTEST, SEIZE, or WOUND, though other categories may appear here as well in certain circumstances when they appear in context with one of these categories, or as other Count categories are added over time. A value of "PROTEST" in this field indicates that this is a count of the number of protesters at a protest.
- **NUMBER**. (integer) This is the actual count being reported. If CountType is "PROTEST" and Number is 126, this means that the source article(s) contained a mention of 126 protesters.
- **OBJECTTYPE**. (character) This records any identifying information as to what the number refers to. For example, a mention of "20 Christian missionaries were arrested" will result in "Christian missionaries" being captured here. This field will be blank in cases where no identifying information could be identified.

GEOGRAPHY

The next set of fields capture the location most closely associated with the count 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 via Leetaru (2012). [1] The closest

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

reference to the count mention is then encoded in these fields. It may not always be possible for the system to locate a match for each count, in which case one or more of the fields may be blank.

To find all counts 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). [2]

- **GEO_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 counts by geographic specificity, for example, extracting only those counts 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.
- **GEO_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 counts 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.
- **GEO_COUNTRYCODE**. (character) This is the 2-character FIPS10-4 country code for the location.
- **GEO_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).
- **GEO_LAT**. (numeric) This is the centroid latitude of the landmark for mapping.
- **GEO_LONG**. (numeric) This is the centroid longitude of the landmark for mapping.
- **GEO_FEATUREID**. (signed integer) This is the GNS or GNIS FeatureID for this location. More information on these values can be found in Leetaru (2012).[3] **NOTE:** This field will be blank or contain a country code, state code, or FIPS code, except when Actor1Geo_Type has a value of 3 or 4. **NOTE**: This field can contain both positive and negative numbers, see Leetaru (2012) for more information on this.

LINKED EVENTS AND SOURCE INFORMATION

The final fields capture any linked events from the primary GDELT event stream and source information for this count.

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

- **CAMEOEVENTIDS**. (character) This field contains a comma-separated list of GlobalEventIDs from the master GDELT event stream of events that were found in the same article(s) as this count was found. This can be used, for example, to associate any protest events from the primary GDELT event database with any "PROTEST" counts found in the same articles to compile a basic measure of how many protesters were involved with the protest. This is useful in that the GDELT event record for the protest encodes many other variables, such as the specific actors involved and their attributes, while the Count record only records that a certain number of protesters were mentioned in an article. The same GlobalEventID may appear in multiple entries if multiple counts were found in the same article mentioning that event or if that event was mentioned in multiple news articles with different counts (reflecting conflicting or evolving information on the details of the event).
- **SOURCES**. This is a semicolon-delimited list of all of the sources publishing articles mentioning this count. For web-based news material, this is the top-level domain the page was from, while for BBC Monitoring service material, "BBC Monitoring" will appear. Thus, it is important to note that this field will contain a mixture of domain names and the phrase "BBC Monitoring".
- **SOURCEURLS**. This is a delimited list of ALL articles mentioning this count. Since URLs can contain a wide variety of characters the phrase "<UDIV>" is used as the delimiter between articles. For web-based news material, this will be the complete URL to the article, while for BBC Monitoring service material, "BBC Monitoring" will appear. Thus, it is important to note that this field will contain a mixture of URLs and the phrase "BBC Monitoring". It is also important to note that this field contains a list of ALL source articles, unlike the primary GDELT event stream, which only lists the first source article mentioning an event in the case an event is mentioned in multiple articles.

DATA FIELDS – GLOBAL KNOWLEDGE GRAPH FILE

This section outlines the specific data fields in the Global Knowledge Graph file (*.gkg.csv). Unlike the primary GDELT event stream, these records are not issued unique identifier numbers, nor are they dated. Users should note that all records are dated with the date of the news media used to identify them, not the actual date of the record, though future releases of the GKG will emphasize adding date information into the file to handle mentions of counts from the past.

The GKG operates around what it calls "namesets", which are essentially a unique pairing of a set of names and other information that appear together in a set of articles. Each morning the GDELT GKG engine processes each news article from the previous day (regardless of whether it contained any GDELT events) and compiles a list of all person names, organization names, locations, overall emotion, any mention of a count of something with respect to a set of predefined indicators (currently eleven categories as of this writing such as arrests and deaths), any mention of a GDELT event, and any mention of a predefined catalog of themes (over 150 themes as of this writing) within that article. It then groups all articles together that contained that same identical set of people, organizations, locations, counts, events, and themes listed in the article (the articles may be about entirely different topics and may have very different emotional scores, as long as they contain the same list of names, locations, counts, and themes). This unique pairing of a set of person names, organization names, locations, counts, and themes is called a "nameset". Thus, an article mentioning Barack Obama, John Kerry, Vladimir Putin, Russia, and Negotiations will result in a unique nameset of "BarackObama-JohnKerry-VladimirPutin-Russia-Negotiations" (the actual technical "unique key" is slightly more complex than this and avoids character collisions). The final output format of the GKG Graph file is each unique nameset per line, with a list of all of the articles containing that nameset from that day.

One consideration to always keep in mind when working with the GKG is that a nameset indicates ONLY that a given set of counts, themes, locations, persons, and organizations appeared together with each other in a given set of articles. No relatedness is implied or suggested by those names appearing together, but those relationships that surface over multiple namesets can traditionally be inferred to imply a certain type of semantic or structural relatedness. For example, a given article might yield John Kerry, Vladimir Putin, and Hassan Rouhani in the Persons field, with United States, Russia, and Iran in the Locations field, and Sanctions in the Themes field. This obviously does not indicate that John Kerry is from Russia or Iran, or that the United States is undergoing Sanctions, but rather indicates that John Kerry is somehow associated with those two locations and that theme through some set of activities. When constructing relationship networks from this data, it is often most valuable to filter the data to exclude names which only appear in a few documents, and to discard connections among names which appear together in only a few documents. For example, John Kerry might become connected to Estonian Foreign Minister Urmas Paet on a given day due to a visit between the two dignitaries, or to Hassan Rouhani over a given month due to active dialogs between the two nations, but when looking over a large period of time, he is most closely related to Barack Obama and the United States. Thus, looking over short periods of time at the connections around John Kerry will yield the important diplomatic and emerging connections of that period, including brief state visits and short exchanges, while extending to longer and longer time periods will surface the longer-term priorities and affiliations of a newsmaker.

Monitoring for change over time is often extremely useful. For example, the United States is often highly associated with discussions of Sanctions since it often leads the call for sanctions against a given nation. Thus, simply constructing a map or list each day of those locations most closely associated with

sanctions will frequently yield the United States at the top of the list. Instead, looking for change over time will cause the United States to largely drop from the list, since its rank rarely changes, while other nations will experience dramatic changes in their ranking on the list as discussion of sanctions increases or decreases with respect to them. It is also important to keep in mind that a Theme merely indicates discussion around a topic, not necessarily action on that topic. For example, Iran experienced a vast surge in association with Sanctions during Hassan Rouhani's visit to the United Nations in September 2013 as the potential for thawing relations between Iran and the United States generated considerable discussion about whether sanctions against Iran would be eased. This considerable increase in media attention to the economic sanctions against Iran merely reflected a strong increase in interest and discussion, but not actual action at that time. Thus, it is important to understand what the presence of a given Theme with respect to a Person or Location indicates in terms of meaning. Themes should be thought of as what the core topic(s) of discussion are around a given entity, rather than an indication of action around that entity.

Linkages amongst Persons, Organizations, and Locations may also reflect the nature of the news environment. For example, if a certain reporter tends to cover a certain company or political leader more often than others, he or she may appear in the graph as tightly connected to that leader, while public relations staff and media spokespersons are often highly prominent in the Persons fields from articles about their companies, since they are often the ones quoted or cited by reporters about that company. Thus, it is common to find a university professor closely associated with a terror organization or a reporter closely associated with a foreign leader or industry, reflecting that these people are regularly in the news commenting or reporting on those organizations, leaders, and industries. One has to always remember that the GKG is based on co-occurrences of names, not on a deeper structural understanding of their organizational relatedness. In this way, GKG reflects how names are contextualized in the news media, rather than capturing the equivalent of an "org chart". This will often result in strange connections among names that might not seem to bear any relationship to one another but is actually capturing how the news media is discussing those individuals.

In this way, the GKG can be used as a "find an expert" resource to locate specific classes of people most closely associated with specific organizations, topics, locations, or leaders. An estimate of the organizational, locational, and functional affiliations of a person can be created by generating a histogram of the values in those fields for all mentions of that person in the GKG. For example, when a histogram of all organization names and TAX_FNCACT themes affiliated with a university faculty member is computed, the university that faculty member belongs to will likely be the most common match, while TAX_FNCACT_FACULTY or TAX_FNCACT_PROFESSOR will likely be the most common TAX_FNCACT values associated with that person's name. Similarly, a reporter will likely have TAX_FNCACT_REPORTER as the most common TAX_FNCACT value associated with his or her name. This can be used to rapidly filter networks constructed from the GKG to identify these kinds of potential expert information sources.

News media often attempt to provide context for the events in their stories by providing background information that explains why the event is of significance and its backstory. When the young Pakistani girl Malala Yousafzai, who had been shot by the Taliban in 2012, was offered citizenship by Canada on October 16, 2013, news coverage recounted others whom Canada had made similar offers to. Reporters focused on the good company she was in, noting that Raoul Wallenberg, Nelson Mandela, the Dalai Lama, Aung San Suu Kyi and the Aga Khan had all been made similar offers. Thus, on that day, the GKG would show connections between Malala and these five international luminaries, reflecting not

necessarily that she had visited or even had contact with those individuals, but rather that the news media was discussing her in their context.

It is important to recognize that the output of the GKG is a direct reflection of what is in the news. News coverage of the world from emerging events to well-known organizations and leaders may not always be accurate or capture the entirety of the available information environment. Like the primary GDELT event stream, the GKG performs extremely sophisticated geographic disambiguation via Leetaru (2012), [4] however it does NOT perform person name disambiguation. Determining whether two different articles about a "John Smith" refer to the same person or to different people both named John Smith is an extremely complex topic. There is an entire field of literature that revolves around this process and a vast array of algorithms for performing disambiguation. One simple approach that can work reasonably well is to keep a running tally for each name as to how many days the name has appeared in the news. Once a name has appeared on a certain number of days, compute a histogram of the locations, organizations, and other names coinciding with that name and associate the name with the most common location, organization, and functional actor (via the Functional Actor theme) that cooccurs with it. In this way, references to President Barack Obama will be most closely associated with the United States Government, the United States, and the functional actor of President. It may be that some mentions of Barack Obama in the news refer to an individual with the same name somewhere else in the world, but it is likely safe to assume that most mentions are to the President of the United States. This is vastly more complicated for individuals who are not heads of state or for whom there is a much greater mix of information and it is highly recommended that users who require highly precise name disambiguation and resolution consult the latest literature and algorithms from the field.

One final important note when using the GKG is that in the digital era news articles can change after they have been published online, including substantial additions, removals, and edits of the article text. While rare, it does occur that an article changes dramatically after we have processed it and thus people, organizations, locations, themes, and events that we record as having been found in that article may no longer be there when the article is accessed at a later date. Events appearing in multiple articles from different sources should still be present in at least one of the cited articles and thus when tracing back additional details of entities, themes, or events, users may on occasion have to read through several of the cited articles if substantial change has occurred in one or more of the source articles. This most often occurs with major breaking stories where substantial information is unknown when the article is first published and the news outlet updates the original story with additional information over the coming hours.

PRIMARY ATTRIBUTES

These fields capture the primary attributes of the nameset.

- **DATE**. (integer) This is the date in YYYYMMDD format on which the news media used to construct this GKG file was published. NOTE that unlike the main GDELT event stream files, this date is the date of publication of the news media from which the information was extracted – if the article discusses events in the past, the date is NOT time-shifted as it is for the GDELT event stream. This date will be the same for all rows in a file and is redundant from a data processing

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

standpoint, but is provided to make it easier to load GKG files directly into an SQL database for analysis.

- **NUMARTS**. (integer) This is the total number of source documents containing one or more mentions of this nameset. This can be used as a method of assessing the "importance" of a nameset: the more discussion of that nameset, 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 namesets during the time period of interest.
- **COUNTS**. (semicolon-delimited blocks, with pound symbol ("#") delimited fields) This is the list of Counts found in this nameset. Each Count found is separated with a semicolon, while the fields within a Count are separated by the pound symbol ("#"). See the documentation for the GKG Count File files to see what each field captures. This captures all of the information found in the Count File and thus when using the GKG Graph file there is no need to separately download the Count File for the same day (the Count File is produced for those who only want to process counts and do not need all of the additional power of the full GKG Graph file).
- **THEMES**. (semi-colon-delimited) This is the list of all Themes found in the nameset. For the complete list of possible themes, see the Category List spreadsheet. At the time of this writing there are over 150 themes currently recognized by the system.
- **LOCATIONS**. (semicolon-delimited blocks, with pound symbol ("#") delimited fields) This is a list of all locations found in the text, extracted through the Leetaru (2012) algorithm. [5] The algorithm is run in a more aggressive stance here than ordinary in order to extract every possible locative referent, so may have a slightly elevated level of false positives.
- **PERSONS**. (semicolon-delimited) This is the list of all person names found in the text, extracted through the Leetaru (2012) algorithm. [6] This name recognition algorithm is unique in that it is specially designed to recognize the African, Asian, and Middle Eastern names that yield significantly reduced accuracy with most name recognition engines.
- **ORGANIZATIONS**. (semicolon-delimited) This is the list of all company and organization names found in the text, extracted through the Leetaru (2012) algorithm. [7] This is a combination of corporations, IGOs, NGOs, and any other local organizations such as a local fair or council. This engine is highly adaptive and is currently tuned to err on the side of inclusion when it is less confident about a match to ensure maximal recall of smaller organizations around the world that are of especial interest to many users of the GKG. Conversely, certain smaller companies with names and contexts that do not provide a sufficient recognition latch may be missed or occasionally misclassified as a person name depending on context. It is highly recommended that users of the Persons and Organizations fields histogram the results and discard names appearing just once or twice to eliminate most of these false positive matches.

EMOTION

At this time, only a single emotional dimension is captured (positive/negative), but several contextual variables around it are also recorded.

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

- **TONE.** (comma-delimited floating point numbers) This field contains a comma-delimited list of six core emotional dimensions, described in more detail below. Each is recorded as a single precision floating point number.
    - **Tone**. This is the average "tone" of all documents containing one or more mentions of this nameset. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This is calculated as Positive Score minus Negative Score. Note that both Positive Score and Negative Score are available separately below as well. A document with a Tone score close to zero may either have low emotional response or may have a Positive Score and Negative Score that are roughly equivalent to each other, such that they nullify each other. These situations can be detected either through looking directly at the Positive Score and Negative Score variables or through the Polarity variable.
    - **Positive Score**. This is the percentage of all words in the article that were found to have a positive emotional connotation. Ranges from 0 to +100.
    - **Negative Score**. This is the percentage of all words in the article that were found to have a negative emotional connotation. Ranges from 0 to +100.
    - **Polarity**. This is the percentage of words that had matches in the tonal dictionary as an indicator of how emotionally polarized or charged the text is. If Polarity is high, but Tone is neutral, this suggests the text was highly emotionally charged, but had roughly equivalent numbers of positively and negatively charged emotional words.
    - **Activity Reference Density**. This is the percentage of words that were active words offering a very basic proxy of the overall "activeness" of the text compared with a clinically descriptive text.
    - **Self/Group Reference Density**. This is the percentage of all words in the article that are pronouns, capturing a combination of self-references and group-based discourse. News media material tends to have very low densities of such language, but this can be used to distinguish certain classes of news media and certain contexts.

LINKED EVENTS AND SOURCE INFORMATION

The final fields capture any linked events from the primary GDELT event stream and source information for this count.

- **CAMEOEVENTIDS**. (character) This field contains a comma-separated list of GlobalEventIDs from the master GDELT event stream of events that were found in the same article(s) as this nameset was found.
- **SOURCES**. This is a semicolon-delimited list of all of the sources publishing articles mentioning this nameset. For web-based news material, this is the top-level domain the page was from, while for BBC Monitoring service material, "BBC Monitoring" will appear. Thus, it is important to note that this field will contain a mixture of domain names and the phrase "BBC Monitoring".
- **SOURCEURLS**. This is a delimited list of ALL articles mentioning this nameset. Since URLs can contain a wide variety of characters the phrase "<UDIV>" is used as the delimiter between articles. For web-based news material, this will be the complete URL to the article, while for BBC Monitoring service material, "BBC Monitoring" will appear. Thus, it is important to note that this field will contain a mixture of URLs and the phrase "BBC Monitoring". It is also important to note that this field contains a list of ALL source articles, unlike the primary GDELT event stream, which only lists the first source article mentioning an event in the case an event is mentioned in multiple articles.