THE GDELT GLOBAL KNOWLEDGE GRAPH (GKG)
DATA FORMAT CODEBOOK V2.1
2/19/2015
http://gdeltproject.org/

INTRODUCTION

This codebook introduces the GDELT Global Knowledge Graph (GKG) Version 2.1, 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. It applies an array of highly sophisticated natural language processing algorithms to each document to compute a range of codified metadata encoding key latent and contextual dimensions of the document. 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.

It has been just short of sixteen months since the original prototype introduction of the GKG 1.0 system on November 3, 2013 and in those fourteen months the GKG system has found application in an incredible number and diversity of fields. The uniqueness of the GKG indicators in capturing the latent dimensions of society that precede physical unrest and their global scope has enabled truly unimaginable new applications. We've learned a lot over the past year in terms of the features and capabilities of greatest interest to the GKG community, and with this Version 2.1 release of the GKG, we are both integrating those new features and moving the GKG into production status (from its original alpha status) in recognition of the widespread production use of the system today.

Due to the vast number of use cases articulated for the GKG, a decision was made at its release 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. 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. Encoding the GKG in XML, JSON, RDF, or other file formats significantly increases the on-disk footprint of the format due to its complexity and size (thus why the GKG is only available in CSV format), though users requiring access to the GKG in these formats can easily write a PERL or Python or similar script to translate the GKG format to any file format needed. The GKG is optimized for fast scanning, storing one record per line and using a tab-delimited format to separate the fields. This makes it possible to use highly optimized fully parallelized streamed parsing to rapidly process the GKG. Similar to the 1.0 format, the files have a ".csv" ending, despite being tab-delimited, to address issues with some software packages that cannot handle ".txt" or ".tsv" endings for parsing tasks.

The new GKG format preserves most of the previous fields in their existing format for backwards compatibility (and we will continue to generate the daily Version 1.0 files in parallel into the future), but

adds a series of new capabilities that greatly enhance what can be done with the GKG data, opening entirely new analytic opportunities. Some of the most significant changes:

- **Realtime Measurement of 2,300 Emotions and Themes**. The GDELT Global Content Analysis Measures (GCAM) module represents what we believe is the largest deployment of sentiment analysis in the world: bringing together 24 emotional measurement packages that together assess more than 2,300 emotions and themes from every article in realtime, multilingual dimensions natively assessing the emotions of 15 languages (Arabic, Basque, Catalan, Chinese, French, Galician, German, Hindi, Indonesian, Korean, Pashto, Portuguese, Russian, Spanish, and Urdu). GCAM is designed to enable unparalleled assessment of the emotional undercurrents and reaction at a planetary scale by bringing together an incredible array of dimensions, from LIWC's "Anxiety" to Lexicoder's "Positivity" to WordNet Affect's "Smugness" to RID's "Passivity".
- **Realtime Translation of 65 Languages**. GDELT 2.0 brings with it the public debut of GDELT Translingual, representing what we believe is the largest realtime streaming news machine translation deployment in the world: all global news that GDELT monitors in 65 languages, representing 98.4% of its daily non-English monitoring volume, is translated in realtime into English for processing through the entire GDELT Event and GKG/GCAM pipelines. GDELT Translingual is designed to allow GDELT to monitor the entire planet at full volume, creating the very first glimpses of a world without language barriers. The GKG system now processes every news report monitored by GDELT across these 65 languages, making it possible to trace people, organizations, locations, themes, and emotions across languages and media systems.
- **Relevant Imagery, Videos, and Social Embeds**. A large fraction of the world's news outlets now specify a hand-selected image for each article to appear when it is shared via social media that represents the core focus of the article. GDELT identifies this imagery in a wide array of formats including Open Graph, Twitter Cards, Google+, IMAGE_SRC, and SailThru formats. In addition, GDELT also uses a set of highly specialized algorithms to analyze the article content itself to identify inline imagery of high likely relevance to the story, along with videos and embedded social media posts (such as embedded Tweets or YouTube or Vine videos), a list of which is compiled. This makes it possible to gain a unique ground-level view into emerging situations anywhere in the world, even in those areas with very little social media penetration, and to act as a kind of curated list of social posts in those areas with strong social use.
- **Quotes, Names, and Amounts**. The world's news contains a wealth of information on food prices, aid promises, numbers of troops, tanks, and protesters, and nearly any other countable item. GDELT 2.0 now attempts to compile a list of all "amounts" expressed in each article to offer numeric context to global events. In parallel, a new Names engine augments the existing Person and Organization names engines by identifying an array of other kinds of proper names, such as named events (Orange Revolution / Umbrella Movement), occurrences like the World Cup, named dates like Holocaust Remembrance Day, on through named legislation like Iran Nuclear Weapon Free Act, Affordable Care Act and Rouge National Urban Park Initiative. Finally, GDELT also identifies attributable quotes from each article, making it possible to see the evolving language used by political leadership across the world.
- **Date Mentions**. We've heard from many of you the desire to encode the list of date references found in news articles and documents in order to identify repeating mentions of specific dates as possible "anniversary violence" indicators. All day, month, and year dates are now extracted from each document.
- **Proximity Context**. Perhaps the greatest change to the overall format from version 1.0 is the introduction of the new Proximity Context capability. The GKG records an enormously rich array

of contextual details from the news, encoding not only the people, organizations, locations and events driving the news, but also functional roles and underlying thematic context. However, with the previous GKG system it was difficult to associate those various data points together. For example, an article might record that Barack Obama, John Kerry, and Vladimir Putin all appeared somewhere in an article together and that the United States and Russia appeared in that article and that the roles of President and Secretary of State were mentioned in that article, but there was no way to associate each person with the corresponding location and functional roles. GKG 2.1 addresses this by providing the approximate character offset of each reference to an object in the original article. While not allowing for deeper semantic association, this new field allows for simple proximity-based contextualization. In the case of the example article above, the mention of United States likely occurs much closer to Barack Obama and John Kerry than to Vladimir Putin, while Secretary of State likely occurs much closer to John Kerry than to the others. In this way, critical information on role, geographic, thematic association, and other connectivity can be explored. Pilot tests have already demonstrated that these proximity indicators can be highly effective at recovering these kinds of functional, thematic, and geographic affiliations.

- **Over 100 New GKG Themes**. There are more than 100 new themes in the GDELT Global Knowledge Graph, ranging from economic indicators like price gouging and the price of heating oil to infrastructure topics like the construction of new power generation capacity to social issues like marginalization and burning in effigy. The list of recognized infectious diseases, ethnic groups, and terrorism organizations has been considerably expanded, and more than 600 global humanitarian and development aid organizations have been added, along with global currencies and massive new taxonomies capturing global animals and plants to aid with tracking species migration and poaching.
- **Extensible XML Block**. GDELT has historically relied primarily on mainstream news coverage for its source material. Whether from print, broadcast, or web-based mediums, news coverage across the world is relatively consistent in the kinds of information it captures. As GDELT encodes an ever-increasing range of materials, including academic journal articles and government reports, additional types of information are available to codify. As a first example of this, Leetaru, Perkins and Rewerts (2014) [1] apply the GKG to encode more than 21 billion words of academic literature, including the entire contents of JSTOR, DTIC, CORE, CireSeerX, and the Internet Archive's 1.6 billion PDFs relevant to Africa and the Middle East. Academic literature contains a list of cited references at the bottom of each article that indicate the papers cited within that paper. This citation list is extremely valuable in constructing citation graphs over the literature to better understand trends and experts. Yet, such citation lists are unique to this class of literature and will not be found in ordinary news material and thus it would be cumbersome to add additional fields to the GKG file format to handle each of these kinds of specialized data types. Instead, the GKG now includes a special field called V2EXTRASXML that is XML formatted and includes these kinds of specialized data types that are applicable only to subsets of the collection. Moving forward, this will allow the GKG to encode highly specialized enhanced information from specialized input streams.
- **Unique Record Identifiers**. To bring the GKG in line with the practices of the GDELT Event Database, every GKG record is now assigned a unique identifier. As with the event database, sequential identifiers do not indicate sequential events, but an identifier uniquely identifies a record across the entire collection. The addition of unique record identifiers to the GKG will make it easier to uniquely refer to a particular GKG record.

[1] http://dlib.org/dlib/september14/leetaru/09leetaru.html

- **Single Data File**. Previously there were two separate GKG data files, one containing Counts only and one containing the full GKG file. The original rationale for having two separate files was that users interested only in counts could download a much smaller daily file, but in practice nearly all applications use the full GKG file in order to make use of its thematic and other data fields to contextualize those counts and to tie them into the GDELT Event Database. Thus, we are eliminating the separate counts-only file to simplify the GKG data environment.
- **Production Status**. The GKG has now moved out of Alpha Experimental Release status and into production status. This means that the file format is now stabilized and will not change.

DIFFERENCES FROM GKG 2.0

The GKG 2.0 file format debuted in September 2014 and several special subcollection datasets were released in that format. With the debut of the GKG 2.1 format in February 2015, the format has remained largely the same, but with the addition of several new fields to accommodate a number of significant enhancements to the GKG system. While it was originally intended to release these new features in the GKG 2.0 format through the V2EXTRASXML field, the integral nature of several of these fields, the desire to more closely align some of them with the format used for the Events dataset, and the need to enable structural mapping of several of the fields to a forthcoming new hierarchical representation, necessitated an upgrade to the GKG file format to the new GKG 2.1 format to accommodate these goals. Users will find that code designed for the GKG 2.0 format can be adapted to the GKG 2.1 format with minimal modification. Since the GKG 2.0 format was only used for a handful of special subcollection datasets and never made an appearance for the daily news content, a GKG 2.0 compatibility feed will not be made available and only the GKG 1.0 and GKG 2.1 formats will be supported for news content.

From a conceptual standpoint, two critical differences between the GKG 2.1/2.0 format and the GKG 1.0 revolve around how entries are clustered and the minimum criteria for an article to be included in the GKG stream. Under the GKG 1.0 format, a deduplication process similar to that used for the Event stream was applied to the daily GKG export, grouping together all articles yielding the same GKG metadata. Thus, two articles listing the same set of locations, themes, people, and organizations would be grouped together in a single row with NumArticles holding a value of 2. With the introduction of the new GCAM system that assess more than 2,300 emotions and themes for each article, it became clear that the GKG 1.0 approach would no longer work, since multiple articles yielding the same locations, themes, people, and organizations might use very different language to discuss them, yielding very different GCAM scores. In addition, the introduction of realtime translation into the GDELT architecture necessitated the ability to identify the provenance of metadata at the document level. Thus, GKG 2.1 no longer clusters documents together based on shared metadata – if 20 articles all contain the same list of extracted locations, themes, people, and organizations, they will appear as 20 separate entries in the GKG stream. The daily GKG 1.0 compatibility stream will, however, still continue to perform clustering. In addition to the clustering change, GKG 2.1 also changes the minimum inclusion criteria for an article to appear in the GKG. Under GKG 1.0 and 2.0, an article was required to have at least one successfully identified and geocoded geographic location before it would be included in the GKG output. However, many topics monitored by GDELT, such as cybersecurity, constitutional discourse, and major policy discussions, often do not have strong geographic centering, with many articles not mentioning even a single location. This was excluding a considerable amount of content from the GKG system that is of high relevance to many GDELT user communities. Thus, beginning with GKG 2.1, an article is included in the GKG stream if it includes ANY successfully extracted information, INCLUDING GCAM emotional scores. An article that contains no recognizable geographic mentions, but lists several political leaders,

or mentions an argument over constitutionalism or a forthcoming policy announcement, will now be included in the GKG stream. Similarly, an article that has no recognizable metadata, but does yield GCAM emotional/thematic scores will also be included. When processing GKG 2.1 files, users should therefore be careful not to include any assumptions in their code as to whether an entry has extracted geographic information and should check the contents of this field for mapping or other geographic applications.

EXTRACTED FIELDS

The following section documents each of the fields contained in the GKG 2.1 format. **Note:** the former format had a NUMARTS field – this has been discontinued due to the new format's support of multiple types of source collections beyond just news media and the requisite need to specify a source collection to interpret document identifiers in the new format (as discussed above). Thus, if multiple documents have identical computed metadata, in 1.0 format they would have been clustered together with NumArts used to indicate the multiple entries, while in the 2.1 format each document has a separate entry in the file. Fields prefaced with "V1" indicate they are identical in format and population to the previous GKG format. Those prefaced with "V1.5" mean they are largely similar, but have some changes. Those prefaced with "V2" are new to the format. Each row represents one document codified by the GKG and each row is tab-delimited for its major fields. **Note**: the "V1/V1.5/V2" designations are not included in the header row of the actual GKG output files. **Note:** the ordering of the fields in the file has substantially changed from Version 2.0 to Version 2.1.

- **GKGRECORDID**. (string) Each GKG record is assigned a globally unique identifier. Unlike the EVENT system, which uses semi-sequential numbering to assign numeric IDs to each event record, the GKG system uses a date-oriented serial number. Each GKG record ID takes the form "YYYYMMDDHHMMSS-X" or "YYYYMMDDHHMMSS-TX" in which the first portion of the ID is the full date+time of the 15 minute update batch that this record was created in, followed by a dash, followed by sequential numbering for all GKG records created as part of that update batch. Records originating from a document that was translated by GDELT Translingual will have a capital "T" appearing immediately after the dash to allow filtering of English/non-English material simply by its record identifier. Thus, the fifth GKG record created as part of the update batch generated at 3:30AM on February 3, 2015 would have a GKGRECORDID of "20150203033000-5" and if it was based on a French-language document that was translated, it would have the ID "20150203033000-T5". This ID can be used to uniquely identify this particular record across the entire GKG database. Note that due to the presence of the dash, this field should be treated as a string field and NOT as a numeric field.
- **V2.1DATE**. (integer) This is the date in YYYYMMDDHHMMSS 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 represents the date of publication of the document 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. **NOTE**: for some special collections this value may be 0 indicating that the field is either not applicable or not known for those materials. For example, OCR'd historical document collections may not have robust metadata on publication date. **NOTE**: the GKG 2.0 format still encoded this date in YYYYMMDD format, while under GKG 2.1 it is now in YYYYMMDDHHMMSS format.

- **V2SOURCECOLLECTIONIDENTIFIER**. (integer) This is a numeric identifier that refers to the source collection the document came from and is used to interpret the DocumentIdentifier in the next column. In essence, it specifies how to interpret the DocumentIdentifier 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 DocumentIdentifier 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 DocumentIdentifier contains the textual citation for the document).
    - 3 = CORE (The document originates from the CORE archive and the DocumentIdentifier contains its DOI, suitable for accessing the original document through the CORE website).
    - 4 = DTIC (The document originates from the DTIC archive and the DocumentIdentifier contains its DOI, suitable for accessing the original document through the DTIC website).
    - 5 = JSTOR (The document originates from the JSTOR archive and the DocumentIdentifier 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 DocumentIdentifier 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).
- **V2SOURCECOMMONNAME**. (text) 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 DocumentIdentifier field.
- **V2DOCUMENTIDENTIFIER**. (text) 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 SOURCECOLLECTION is equal to 1, this field will contain a fully-qualified URL suitable for direct access. If SOURCECOLLECTION 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 SOURCECOLLECTION 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.
- **V1COUNTS**. (semicolon-delimited blocks, with pound symbol ("#") delimited fields) This is the list of Counts found in this document. Each Count found is separated with a semicolon, while

the fields within a Count are separated by the pound symbol ("#"). Unlike the primary GDELT event stream, these records are not issued unique identifier numbers, nor are they dated. As an example of how to interpret this file, an entry with CountType=KILL, Number=47, ObjectType="jihadists" indicates that the article stated that 47 jihadists were killed. This field is identical in format and population as the corresponding field in the GKG 1.0 format.

  - **Count Type.** (text) 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 would indicate that this is a count of the number of protesters at a protest.
  - **Count.** (integer) This is the actual count being reported. If CountType is "PROTEST" and Number is 126, this means that the source article contained a mention of 126 protesters.
  - **Object Type.** (text) 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.
  - **Location Type**. See the documentation for V1Locations below.
  - **Location FullName**. See the documentation for V1Locations below.
  - **Location CountryCode**. See the documentation for V1Locations below.
  - **Location ADM1Code.** See the documentation for V1Locations below.
  - **Location Latitude**. See the documentation for V1Locations below.
  - **Location Longitude**. See the documentation for V1Locations below.
  - **Location FeatureID**. See the documentation for V1Locations below.
- **V2.1COUNTS**. (semicolon-delimited blocks, with pound symbol ("#") delimited fields) This field is identical to the V1COUNTS field except that it adds a final additional field to the end of each entry that records its approximate character offset in the document, allowing it to be associated with other entries from other "V2ENHANCED" fields (or Events) that appear in closest proximity to it. **Note:** unlike the other location-related fields, the Counts field does NOT add ADM2 support at this time. This is to maintain compatibility with assumptions that many applications make about the contents of the Count field. Those applications needing ADM2 support for Counts should cross-reference the FeatureID field of a given Count against the V2Locations field to determine its ADM2 value.
- **V1THEMES**. (semi-colon-delimited) This is the list of all Themes found in the document. For the complete list of possible themes, see the Category List spreadsheet. At the time of this writing there are over 275 themes currently recognized by the system. This field is identical in format and population as the corresponding field in the GKG 1.0 format.
- **V2ENHANCEDTHEMES.** (semicolon-delimited blocks, with comma-delimited fields) This contains a list of all GKG themes referenced in the document, along with the character offsets of approximately where in the document they were found. For the complete list of possible themes, see the Category List spreadsheet. At the time of this writing there are over 300 themes currently recognized by the system. Each theme reference is separated by a semicolon, and within each reference, the name of the theme is specified first, followed by a comma, and then the approximate character offset of the reference of that theme in the document, allowing it to be associated with other entries from other "V2ENHANCED" fields that appear in closest

proximity to it. If a theme is mentioned multiple times in a document, each mention will appear separately in this field.

- **V1LOCATIONS**. (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. [2] 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. **NOTE**: some locations have multiple accepted formal or informal names and this field is collapsed on name, rather than feature (since in some applications the understanding of a geographic feature differs based on which name was used to reference it). In cases where it is necessary to collapse by feature, 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). [3] This field is identical in format and population as the corresponding field in the GKG 1.0 format. **NOTE**: there was an error in this field from 2/19/2015 through midday 3/1/2015 that caused the CountryCode field to list the wrong country code in some cases.
    - **Location 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 contain its textual country or ADM1 code instead of a numeric featureid.
    - **Location FullName**. (text) 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.
    - **Location CountryCode**. (text) This is the 2-character FIPS10-4 country code for the location. **Note:** GDELT continues to use the FIPS10-4 codes under USG guidance while GNS continues its formal transition to the successor Geopolitical Entities, Names, and Codes (GENC) Standard (the US Government profile of ISO 3166).
    - **Location ADM1Code**. (text) 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

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

shortform of the state's name (such as "TX" for Texas). **Note:** see the notice above for CountryCode regarding the FIPS10-4 / GENC transition. **Note**: to obtain ADM2 (district-level) assignments for locations, you can either perform a spatial join against a ShapeFile template in any GIS software, or cross-walk the FeatureID to the GNIS/GNS databases – this will provide additional fields such as ADM2 codes and MGRS grid references for GNS.

  - **Location Latitude**. (floating point number) This is the centroid latitude of the landmark for mapping. In the case of a country or administrative division this will reflect the centroid of that entire country/division.
  - **Location Longitude**. (floating point number) This is the centroid longitude of the landmark for mapping. In the case of a country or administrative division this will reflect the centroid of that entire country/division.
  - **Location FeatureID**. (text OR signed integer) This is the numeric GNS or GNIS FeatureID for this location OR a textual country or ADM1 code. More information on these values can be found in Leetaru (2012).[4] **Note:** This field will be blank or contain a textual ADM1 code for country or ADM1-level matches – see above. **Note**: For numeric GNS or GNIS FeatureIDs, this field can contain both positive and negative numbers, see Leetaru (2012) for more information on this.

- **V2ENHANCEDLOCATIONS**. (semicolon-delimited blocks, with pound symbol ("#") delimited fields) This field is identical to the V1LOCATIONS field with the primary exception of an extra field appended to the end of each location block after its FeatureID that lists the approximate character offset of the reference to that location in the text. In addition, if a location appears multiple times in the article, it will be listed multiple times in this field. The only other modification from V1LOCATIONS is the addition of a single new field "Location ADM2Code" in between "Location ADM1Code" and "Location Latitude". [5] **NOTE**: there was an error in this field from 2/19/2015 through midday 3/1/2015 that caused the CountryCode field to list the wrong country code in some cases.
- **V1PERSONS**. (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. This field is identical in format and population as the corresponding field in the GKG 1.0 format.
- **V2ENHANCEDPERSONS**. (semicolon-delimited blocks, with comma-delimited fields) This contains a list of all person names referenced in the document, along with the character offsets of approximately where in the document they were found. Each person reference is separated by a semicolon, and within each reference, the person name is specified first, followed by a comma, and then the approximate character offset of the reference of that person in the document, allowing it to be associated with other entries from other "V2ENHANCED" fields that appear in closest proximity to it. If a person is mentioned multiple times in a document, each mention will appear separately in this field.
- **V1ORGANIZATIONS**. (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

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

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. This field is identical in format and population as the corresponding field in the GKG 1.0 format.

- **V2ENHANCEDORGANIZATIONS**. (semicolon-delimited blocks, with comma-delimited fields) This contains a list of all organizations/companies referenced in the document, along with the character offsets of approximately where in the document they were found. Each organization reference is separated by a semicolon, and within each reference, the name of the organization is specified first, followed by a comma, and then the approximate character offset of the reference of that organization in the document, allowing it to be associated with other entries from other "V2ENHANCED" fields that appear in closest proximity to it. If an organization is mentioned multiple times in a document, each mention will appear separately in this field.
- **V1.5TONE.** (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. This field is nearly identical in format and population as the corresponding field in the GKG 1.0 format with the sole exception of adding the single new WordCount variable at the end.
    - **Tone**. (floating point number) This is the average "tone" of the document as a whole. 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**. (floating point number) 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**. (floating point number) This is the percentage of all words in the article that were found to have a positive emotional connotation. Ranges from 0 to +100.
    - **Polarity**. (floating point number) 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**. (floating point number) 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**. (floating point number) 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.

    - **Word Count**. (integer) This is the total number of words in the document. This field was added in version 1.5 of the format.
- **V2.1ENHANCEDDATES**. (semicolon-delimited blocks, with comma-delimited fields) This contains a list of all date references in the document, along with the character offsets of approximately where in the document they were found. If a date was mentioned multiple times in a document, it will appear multiple times in this field, once for each mention. Each date reference is separated by a semicolon, while the fields within a date are separated by commas. **NOTE**: this field is identical to GKG 2.0 with the sole exception of the addition of one additional Date Resolution type (4 = dates that include a month and day, but not a year).
    - **Date Resolution.** This indicates whether the date was a month-day date that did not specify a year (4), a fully-resolved day-level date that included the year (3), a month-level date that included the year but not a day (2), or a year-level (1) date that did not include month or day-level information.
    - **Month.** This is the month of the date represented as 1-12. For year dates this field will contain a 0.
    - **Day.** This is the day of the date represented as 1-31. For month and year dates this field will contain a 0.
    - **Year.** This is the year of the date. For Resolution=4 dates that include a month and day, but not a year, this field will contain a 0.
    - **Offset.** This is the character offset of the date within the document, indicating approximately where it was found in the body. This can be used to associate the date with the entries from other "V2ENHANCED" fields that appeared in closest proximity to it.
- **V2GCAM**. (comma-delimited blocks, with colon-delimited key/value pairs) The Global Content Analysis Measures (GCAM) system runs an array of content analysis systems over each document and compiles their results into this field. New content analysis systems will be constantly added to the GCAM pipeline over time, meaning the set of available fields will constantly grow over time. Given that the GCAM system is debuting with over 2,300 dimensions and will likely grow to include several thousand more dimensions within the coming months, it differs in its approach to encoding matches from the GKG's native thematic coding system. Instead of displaying the full English name of a content analysis dictionary or dimension, it assigns each dictionary a unique numeric identifier (DictionaryID) and each dimension within that dictionary is assigned a unique identifier from 1 to the number of dimensions in the dictionary (DimensionID). Each dimension of each dictionary is assessed on a document and ONLY those dimensions that had one or more matches onto the document are reported. If a dimension did not register any matches on a document, it is not reported in order to save space. Thus, the absence of a dimension in this field can be interpreted as a score of 0. Each dimension's score is written to the V2GCAM field separated by a comma. For each dimension, a numeric "key" identifies it of the form "DictionaryID.DimensionID", followed by a colon, followed by its score. Most dictionaries are count-based, meaning they report how many words in the document were found in that dictionary. Thus, a score of 18 would mean that 18 words from the document were found in that dictionary. Count-based dimensions have a key that begins with "c". Some dictionaries, such as SentiWordNet and SentiWords actually assign each word a numeric score and the output of the tool is the average of those scores for that document. For those dictionaries, an entry will report the number of words in the document that matched into that dictionary, and a separate entry, beginning with a "v" instead of a "c" will report its floating-point average value. The very first entry in the field has the special reserved key of "wc" and reports the total number of words in the document – this can be used to divide

the score of any word-count field to convert to a percentage density score. As an example, assume a document with 125 words. The General Inquirer dictionary has been assigned the DictionaryID of 2 and its "Bodypt" dimension has a DimensionID of 21. SentiWordNet has a DictionaryID of 10 and its "Positive" dimension has a DimensionID of 1. Thus, the V2GCAM field for a document might look like "wc:125,c2.21:4,c10.1:40,v10.1:3.21111111" indicating that the document had 125 words, that 4 of those words were found the General Inquirer "Bodypt" lexicon, that 40 of those words were found in the SentiWordNet lexicon, and that the average numeric score of all of the words found in the SentiWordNet lexicon was 3.21111111. For a complete list of the available dimensions, along with their assigned DictionaryID and DimensionID codes, their assigned key, and their human name and full citation to cite that dimension, please see the GCAM Master Codebook. [8] **NOTE**: the scores for all dimensions, both English and non-English dimensions, will be listed together in this field – please see the codebooks to determine the source language of a specific dimension. **NOTE**: if non-English dictionaries are available for a given language and generated at least one match for that document, an additional "nwc" entry will be added which reflects the word count in the native language, since languages may have differing word counts in their native and translated forms. This count will be absent if no native dictionaries yielded a match for the document.

- **V2.1SHARINGIMAGE**. (textual URL) Many news websites specify a so-called "sharing image" for each article in which the news outlet manually specifies a particular image to be displayed when that article is shared via social media or other formats. Not all news outlets specify a sharing image and some sites simply use their logo, but for those that do use this field, it represents the outlet's selection of the single image that best captures the overall focus and contents of the story. GDELT currently recognizes a variety of formats for specifying this image, including Open Graph, Twitter Cards, Google+, IMAGE_SRC, and SailThru formats, among others.
- **V2.1RELATEDIMAGES**. (semicolon-delimited list of URLs). News articles frequently include photographs, figures, and other imagery to illustrate the story, ranging from a single illustrative photograph at top, to lengthy photo essays interspersed through the entirety of an article. Such imagery lends a rich visual tapestry to a news report, helping to clarify, for example, whether an article about a "protest blocking a highway" involves hundreds of activists along its length, or just a handful of people in one location, or whether a gas explosion leveled a building or merely shattered its windows. GDELT uses a suite of highly sophisticated algorithms to actually "read" through each article in the way a human would, evaluating each image on to determine its relevance, based on positioning, captioning, referencing, and context, and compiles a list of the URLs of the images it deems to be most relevant to the article. Thus, unrelated inset boxes, advertisements, and other imagery are ignored and this field contains only a list of images most illustrative of the core of the story. This feature is in alpha release and involves a number of highly complex algorithms working together in concert and thus may make mistakes. We will be improving this algorithm over time and would appreciate any feedback you may have on the kinds of images it incorrectly includes and those that it misses.
- **V2.1SOCIALIMAGEEMBEDS**. (semicolon-delimited list of URLs). News websites are increasingly embedding image-based social media posts inline in their articles to illustrate them with realtime reaction or citizen reporting from the ground. GDELT currently recognizes embedded image-based Twitter and Instagram posts and records their URLs in this field. Only those posts containing imagery are included in this field. This acts as a form of social media "curation" in which news outlets are wading through the deluge of social media reaction or reporting about a specific situation and hand-selecting key image posts deemed of greatest relevance,

[8] http://data.gdeltproject.org/documentation/GCAM-MASTERCODEBOOK.xlsx

significance, credibly, and/or interest to their audiences. Only image-based embedded posts are included in this field – videos are identified in the following field.

- **V2.1SOCIALVIDEOEMBEDS**. (semicolon-delimited list of URLs). News websites are increasingly embedding videos inline in their articles to illustrate them with realtime reaction or citizen reporting from the ground. Some news outlets that also have television properties may cross-link their television reporting into their web-based presentation. GDELT currently recognizes YouTube, DailyMotion, Vimeo, and Vine videos embedded in articles and records their URLs in this field. Similarly to the field above, this allows for a form of social media "curation" of the videos deemed by the mainstream media to be of greatest relevance, significance, credibly, and/or interest to their audiences.
- **V2.1QUOTATIONS**. (pound-delimited ("#") blocks, with pipe-delimited ("|") fields). News coverage frequently features excerpted statements from participants in an event and/or those affected by it and these quotations can offer critical insights into differing perspectives and emotions surrounding that event. GDELT identifies and extracts all quoted statements from each article and additionally attempts to identify the verb introducing the quote to help lend additional context, separating "John retorted…" from "John agreed…" to show whether the speaker was agreeing with or rejecting the statement being made. Each quoted statement is separated by a "#" character, and within each block the following fields appear, separated by pipe ("|") symbols:
    - **Offset.** This is the character offset of the quoted statement within the document, indicating approximately where it was found in the body. This can be used to associate the date with the entries from other "V2ENHANCED" fields that appeared in closest proximity to it.
    - **Length.** This is the length of the quoted statement in characters.
    - **Verb.** This is the verb used to introduce the quote, allowing for separation of agreement versus disagreement quotes. May not be present for all quotes and not all verbs are recognized for this field.
    - **Quote.** The actual quotation itself.
- **V2.1ALLNAMES**. (semicolon-delimited blocks, with comma-delimited fields) This field contains a list of all proper names referenced in the document, along with the character offsets of approximately where in the document they were found. Unlike the V2ENHANCEDPERSONS and V2ENHANCEDORGANIZATIONS fields, which are restricted to person and organization names, respectively, this field records ALL proper names referenced in the article, ranging from named events like the Orange Revolution, Umbrella Movement, and Arab Spring, to movements like the Civil Rights Movement, to festivals and occurrences like the Cannes Film Festival and World Cup, to named wars like World War I, to named dates like Martin Luther King Day and Holocaust Remembrance Day, to named legislation like Iran Nuclear Weapon Free Act, Affordable Care Act and Rouge National Urban Park Initiative. This field goes beyond people and organizations to capturing a much broader view of the named events, objects, initiatives, laws, and other types of names in each article. Each name reference is separated by a semicolon, and within each reference, the name is specified first, followed by a comma, and then the approximate character offset of the reference of that name in the document, allowing it to be associated with other entries from other "V2ENHANCED" fields that appear in closest proximity to it. If a name is mentioned multiple times in a document, each mention will appear separately in this field. This field is designed to be maximally inclusive and in cases of ambiguity, to err on the side of inclusion of a name.
- **V2.1AMOUNTS**. (semicolon-delimited blocks, with comma-delimited fields) This field contains a list of all precise numeric amounts referenced in the document, along with the character

offsets of approximately where in the document they were found. Its primary role is to allow for rapid numeric assessment of evolving situations (such as mentions of everything from the number of affected households to the estimated dollar amount of damage to the number of relief trucks and troops being sent into the area, to the price of food and medicine in the affected zone) and general assessment of geographies and topics. Both textual and numeric formats are supported ("twenty-five trucks", "two million displaced civilians", "hundreds of millions of dollars", "$1.25 billion was spent", "75 trucks were dispatched", "1,345 houses were affected", "we spent $25m on it", etc). At this time, percentages are not supported due to the large amount of additional document context required for meaningful deciphering ("reduced by 45%" is meaningless without understanding what was reduced and whether the reduction was good or bad, often requiring looking across the entire enclosing paragraph for context). This field is designed to be maximally inclusive and in cases of ambiguity, to err on the side of inclusion of an amount even if the object of the amount is more difficult to decipher.

  - **Amount**. This is the precise numeric value of the amount. Embedded commas are removed ("1,345,123" becomes 1345123), but decimal numbers are left as is (thus this field can range from a floating point number to a "long long" integer). Numbers in textual or mixed numeric-textual format ("such as "2m" or "two million" or "tens of millions") are converted to numeric digit representation.
  - **Object**. This is the object that the amount is of or refers to. Thus, "20,000 combat soldiers" will result in "20000" in the Amount field and "combat soldiers" in this field.
  - **Offset.** This is the character offset of the quoted statement within the document, indicating approximately where it was found in the body. This can be used to associate the date with the entries from other "V2ENHANCED" fields that appeared in closest proximity to it.

- **V2.1TRANSLATIONINFO**. (semicolon-delimited fields) This field 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".
  - **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 [9] 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"

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

indicates that the Stanford Chinese Word Segmenter [10] 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.

- **V2EXTRASXML**. (special XML formatted) This field is reserved to hold special non-standard data applicable to special subsets of the GDELT collection. It is unique among the other GKG fields in that it is XML-formatted and the specific format of a given block within this field is highly customized. At the time of this writing it currently is used to hold the citations list for the academic journal article subcollection [11] and is blank for news content.
    - **CITEDREFERENCESLIST.** This block holds the results of the parsed cited references list that appeared at the bottom of the article, as extracted by the ParsCit software. [12] The ParsCit system is based on machine learning algorithms which can exhibit a significant amount of error and/or vary by source material. Within this block, each citation is enclosed in a <CITATION></CITATION> block. Within that block appear the following fields. **Note**: the fields may not appear in precisely this order and not all fields may be present for all citations, so parsing of this field should be flexible. For more information on the meaning of each field, please see the documentation for ParsCit. [13] This block is only available for the academic journal article subcollection.
        - **Authors**. This is a nested block with an outer set of tags of <AUTHORS> </AUTHORS> containing one or more inner blocks of <AUTHOR></AUTHOR>. Each inner block contains the name of an author of the cited paper. If a paper has multiple authors, there will be an <AUTHOR></AUTHOR> inner block for each author. Author names are order-standardized ("Leetaru, Kalev Hannes" will be normalized to "Kalev Hannes Leetaru") but are not otherwise normalized and thus "K Leetaru", "Kalev Leetaru", "Kalev H. Leetaru" and "Kalev Hannes Leetaru" would all appear as distinct author entries. Applications requiring name disambiguation will need to perform that task themselves.
        - **Title**. This is the title of the cited work if it is an article.
        - **BookTitle**. This is the title of the cited work if it is a book.
        - **Date**. This is the date of the cited work.
        - **Journal**. The journal the cited work was published in.
        - **Volume**. The volume of the journal issue the cited work was published in.
        - **Issue**. The issue of the journal issue the cited work was published in.
        - **Pages**. This is the page range of the cited work.
        - **Institution**. This is the institutional affiliation of the cited work.
        - **Publisher**. The publisher of the cited work.
        - **Location**. The location of the publisher of the cited work.
        - **Marker**. This is the textual marker used to identify the work in the text (such as "Leetaru et al, 2014"). This can be used if you have access to the original article to locate references to the cited work in the article.

[10] http://nlp.stanford.edu/software/segmenter.shtml
[11] http://blog.gdeltproject.org/announcing-the-africa-and-middle-east-global-academic-literature-knowledge-graph-ame-gkg/
[12] http://aye.comp.nus.edu.sg/parsCit/
[13] http://aye.comp.nus.edu.sg/parsCit/