GReVD Coding rules v0.92

Compiling violent deaths from media coding sources (Primary sources: ACLED, GTD, UCDP)
April 1, 2020

The overarching objective of the GReVD initiative is a common recording platform for every violent death by location and time. Because of the scale of the initiative, it will be necessary to draw data on violent deaths (of all types) from multiple sources. This is an ambitious goal that will not be fully realized immediately, but nevertheless, it must begin somewhere. This document describes an initial set of coding rules for extracting records of violent deaths from the three leading global datasets that collect conflict-related deaths (ACLED[1], GTD[2], and UCDP[3]).

All three primary data sources are compiled by members of the GReVD consortium and every effort will be made in compiling GReVD to draw on the most recent data from these sources and provide proper attribution of original coding from these sources. GReVD differs from current coding processes of ACLED, GTD and UCDP as it is not event based like these datasets - every entry (row) in the Registry will be a single fatality. As a result, GReVD is not a competitor dataset to partner initiatives, rather it is an attempt to build consensus by GReVD members for use by policymakers and practitioners to answer the vital question of ***how many violent deaths there are globally***, using rigorous standards, a single definition and a unified recording process.

It is expected that future versions of the Registry will include coding from other sources, including administrative sources, following protocols based on these being developed below.[4] This document is a draft and these protocols are pilots – *data produced using these protocols is draft and for research purposes only, it should not be used as an estimate or count* (for a variety of reasons, see the Gaps Report).

The coding process is broken into two stages described below (Stage A (ID, creating a row for every violent death from each of the three source datasets) and Stage B (organizing data available for each entry from multiple sources)). The variables described below are divided into five main categories[5] (A1. ID (4 variables), B1. Date (5), B2. Location (4), B3. Tags (1), and B4. Source (6)). Within these five categories, each Registry entry will have a total of 20 variables (columns). It is not expected (or necessary) that all fields will have data at this stage for each entry.

STAGE A (ID, 4 Variables)

Stage A creates a row for every fatality event identified in three partners (ACLED, GTD, UCDP) and associates an ID variable to each entry in the GReVD registry.[6] The unique 22-digit identifier (grevd_id) is constructed from three variables (source, source id and count). It is a principle of

[1] Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. (2010). "Introducing ACLED-Armed Conflict Location and Event Data." Journal of Peace Research 47(5) 651-660.

[2] START (National Consortium for the Study of Terrorism and Responses to Terrorism). 2013. Global Terrorism Database. Accessed December 4, 2019. http://www.start.umd.edu/gtd.

[3] Sundberg, Ralph, and Erik Melander, 2013, "Introducing the UCDP Georeferenced Event Dataset", Journal of Peace Research, vol.50, no.4, 523-532

[4] Fatality data produced by ACLED, GTD and UCDP, coded with location and time, is, in many cases, the best data available on violent deaths. Standards will be set based on consortium members' good practices to promote better quality data from other sources, including, eventually, administrative data.

[5] These categories are based on initial work comparing these three datasets (see Conflict_datasets_review.xlsx directory and a supplementary Data_annotation.docx file, available from the consortium).

[6] As a result, there is no fatality count in the Registry, the count is the Registry. Later versions of the registry will allow users to filter by means, motive, precision, etc. and thereby define violent death.

GReVD that we trace every entry to original source dataset, to ensure proper attribution to GReVD data providers and, later, for validation processes. Please note that NO efforts are made (yet) to reconcile overlapping/duplicate entries from different sources, therefore, this dataset CANNOT be used for counting violent deaths. We use fictional examples below for demonstration only.

A1) Extract coding information from primary sources for each entry

The .do file for Stage A is titled grevd_stage_a.do and is available within the consortium, on request.

1. *eventid_fatality_count*

description: For each event (identified by original source), create a single entry for each fatality, with a cumulative count from 1 to N, where N = total fatalities attached to the event by the original coding source (original variable name for each source is described in *italics*, under instructions, below). Define new variable *eventid_fatality_count.* This is a six-digit integer variable (1 to 999 999). Only events that have associated fatalities will have line entries, so there will be no missing values.

ACLED instructions: Cumulative count, coded from *fatalities*>0[7]

GTD Instructions: Cumulative count coded from *nkill>0,* ignore if *nkill* is blank/missing value[8]

UCDP Instructions: Cumulative count coded from *best*

Example: an event that resulted in 3 fatalities will have three rows in the Registry under the new variable/column *eventid_fatality_count* :

eventid_fatality_count
000001
000002
000003

2. *eventid_source_no*

description: Assign a source number for each original source ACLED=1001; GTD=1002, UCDP=1003 (new variable name: *eventid_source_no).* This is a four-digit numerical categorical variable (1001 to 9999), allowing for up to 8998 sources. Future versions of the Registry will also code from other sources (e.g., intentional homicide data from UNODC), these will be numbered sequentially when they are added.

Example (continued from previous): if ACLED recorded an event resulting in 3 fatalities, GTD reported an incident with one fatality and UCDP reported a battle with 2 fatalities, then the full dataset would include six entries with *eventid_fatality_count and eventid_source_no* as follows:

eventid_fatality_count	eventid_source_no
000001	1001
000002	1001
000003	1001

[7] “ACLED does not have a minimum fatality level for inclusion. If reports say "several", "many", or "few" fatalities and the exact number is unknown, then "10" is recorded as the total. If "dozen" then "12" is recorded. "100" is recorded if report mentions "hundreds" or "massacre". If there is no reference made to fatalities, then "0" is recorded”. See ACLED Codebook (2019).

[8] “If figure is not reported or too vague, this field is left blank”. See GTD Codebook (2019)

000001	1002
000001	1003
000002	1003

3. *eventid_original*

description: for each fatality, use the event identification from original source, coded as *eventid_original.* This is a numeric variable from source (ACLED, GTD, UCDP). Format is integer, based on original source - ACLED (7-digit), GTD (12-digit), UCDP (1 to 6 digits).

ACLED instructions: Code on *data_id*[9]

GTD Instructions: Code on *eventid*

UCDP Instructions: Code on *id*

Example (continued from previous examples), using random event IDs:

eventid_fatality_count	eventid_source_no	eventid_original
000001	1001	000004745705
000002	1001	000004745705
000003	1001	000004745705
000001	1002	201612280045
000001	1003	000000233552
000002	1003	000000233552

4. grevd_id

description: create a composite unique identifier for GReVD by concatenation of variables, in order shown below. Thus a *grevd_id* will be 22 integers: XXXXXXYYYYYYYYYYYYZZZZZZ where Xs are six digits using *eventid_source_no*, Ys are 12 digits from *eventid_original* (adding zeroes as necessary) and Zs are six digits from *eventid_fatality_count*.

Example: eventually, a *grevd_id* is created building on the previous three steps (note order).

eventid_source_no	eventid_original	eventid_fatality_count	grevd_id
1001	000004745705	000001	1001000004745705000001
1001	000004745705	000002	1001000004745705000002
1001	000004745705	000003	1001000004745705000003
1002	201612280045	000001	1002201612280045000001
1003	000000233552	000001	1003000000233552000001
1003	000000233552	000002	1003000000233552000002

STAGE B (Importing additional data for each fatality)

With the four variables above, every fatality coded by ACLED, GTD and UCDP is uniquely defined (within each data source, but has still not been deduplicated between data sources) and can be

[9] In ACLED, a variable '*event_id_no_cnty*' comes closest to how GTD and UCDP treat their event ID variables. Since it turns out to have duplicates and therefore not suitable for constructing the *grevd_id*, a variable *data_id*, which is a unique identifier, is used instead.

traced back to the original source data. In Stage B, any additional data available on the fatality is coded in the Registry from the primary sources.

The .do file for Stage B is titled grevd_stage_ab.do

B1. DATE variables (5 Variables)

Optimally, each Registry entry would be coded with perfect temporal accuracy (not just to the day, but to the hour and minute – this remains aspirational) to reduce duplication. Currently, data is commonly available for year, and often by month and date, for event data.

In some cases, a range of dates is given for an event. We remain agnostic about which date should be assigned to each fatality when a range is given. Also, we don't want to lose information from original coding for source datasets. In the event that a date range is given, a midpoint is selected and then a *width_range* is defined based on the range.

To reduce loss of information, but maintain parsimony in coding, we construct a *temporal_precision* variable, based on precision variables coded from primary sources, but with some generalization between data sources.

5. *year*

Integer, four-digits.
Rule for ACLED: Code on *event_date YYYY* (DD-MM-YYYY format)
Rule for GTD: Code on *iyear* [10]
Rule for UCDP: code on midpoint for *date_start* and *date_end* variables (YYYY-MM-DD format)[11]

6. *month*

Integer, two-digits.
Rule for ACLED: Code on *event_date, MM* (DD-MM-YYYY format)
Rule for GTD: Code on *imonth*[12]
Rule for UCDP: code on midpoint for *date_start* and *date_end* variables (YYYY-MM-DD format)

7. *day*

Integer, two-digits.
Rule for ACLED: Code on *event_date, DD* (DD-MM-YYYY format)
Rule for GTD: Code on *iday*[13]
Rule for UCDP: code on midpoint for *date_start* and *date_end* variables (YYYY-MM-DD format)

8. *width_range*

This variable is in days (Integer, three-digits). It is defined based on the range of possible dates reported for those events that took place over more than one day.

Rule for ACLED: *width_range* does not apply to ACLED events because every event is split into multiple days. To give an example from the ACLED Codebook (2019) and data downloaded

[10] In GTD, if incident(s) occurred over an extended period, the year when the incident was initiated is recorded. See GTD Codebook (2019).

[11] The CURRENT .do file codes on "*date start*" – need to confirm with UCDP that midpoint is reasonable – see also width_range below.

[12] GTD registers "0" if the exact month of the event is unknown during 1970-2011; For attacks after 2011, the midpoint of the range of possible dates reported in sources is taken and the full range is recorded in the approxdate. See GTD Codebook (2019).

[13] if incident(s) occurred over an extended period, the day when the incident was initiated is recorded; for unknown day the same rule as for imonth applies. See GTD Codebook (2019).

through ACLED's Data Export Tool webpage, those events that occurred across several days (e.g., March 1, 1999 – March 5, 1999) with violent activities reported each day are coded as five different events with a different date for each entry.[14]

Therefore, there is no width_range for fatalities sourced from ACLED (*width_range=0*).

Rule for GTD: code on *iday, imonth, iyear*[15]*, extended* and *resolution* (YYYY-MM-DD format) variables. Define *width_range* based on the range of days if it is an extended incident[16] (*extended=1*) and if the date recorded in *resolution* is different from *iday, imonth, iyear.*

Code *width_range =0* if *extended =* 0.

Rule for UCDP: code on midpoint for *date_start* and *date_end* variables. Define *width_range* based on the range of days if *date_end* has a different value (>, is higher) than *date_start.*

Code *width_range = 0* if *date_start = date_end.*

9. *temporal_precision*

ACLED instructions: Code on *time_precision*

GTD instructions: GTD does not record the temporal precision for each incident. However, GTD records an approximate date (*approxdate*) for incidents where exact day and/or month is unknown. Due to different coding rules introduced in GTD for incidents during the period of 1970-2011 and for the incidents after 2011 we adopt the respective rules in our coding for both time periods (1970-2011 and post-2011) [17]. Therefore, *approxdate* is used to code *temporal_precision* using *iyear, imonth* and *iday* as follows (also see table 1 below):

This is a categorical variable (define **1-** if *approxdate* is missing, ie. source material reports the exact date of the attack, therefore, *imonth* and *iday* have a value; **2-** (post-2011 events): if both *iday* and *approxdate* have a value; **3-** (1970–2011 events): if *iday* equals 0 and *imonth* has a value (*imonth*>0); (post-2011 events): if both *imonth* and *approxdate* have a value, but *iday* is missing/0; **4-** (1970–2011 events): if *imonth* equals 0.

UCDP instructions: Code on *date_prec*

The table below lists each value of *temporal_precision* in GReVD based on precision variables coded from primary sources and some alignment between data sources.

[14] An alternative, to be confirmed with ACLED, is identified in the API User Guide which indicates that there are two more variables available to access when downloading through ACLED API calls: *first_event_date* and *last_event _date* (YYYY-MM-DD, the dates that the first and the last events occurred for a particular actor type). In the event that it is relevant and becomes possible to generate these two variables, we will then be able to define *width_range* for ACLED events in the Registry, if necessary.

[15] Incidents recorded in GTD can go over years (e.g, if an attack was initiated at the end of the year), as well as over months and days. Most of such types of attack are incidents of 'Hostage Taking (Kidnapping)'. Even though not all of them result in fatalities, there are still events that record at least one fatality. A few examples from GTD are an attack in Colombia resulting in one fatality over the period of 31.12.1981 – 08.01.1982 and a hostage taking (kidnapping) in Syria with two fatalities over 11.01.2016 – 10.03.2016.

[16] GTD's definition of extended incident: 1="yes", if duration of incident is more than 24 hours; 0="No", if duration of incident is less than 24 hours.

[17] For attacks that happened after 2011 if the exact day of the event is unknown, GTD records the midpoint of the range of possible dates reported in source materials in *iday/imonth* and the full range is recorded in the Approximate Date (*approxdate*). For attacks that took place between 1970 and 2011 GTD codes "0" if the exact day and/or month is unknown. See GTD Codebook (2019).

Table 1 Description of the coding values in GReVD and the temporal precision scale across three sources.

GReVD's coding values	**ACLED** (what *time_precision* values mean)	**GTD** (a new variable coding on *approxdate; iyear; imonth; iday*)	**UCDP** (what *data_prec* values mean)
1	**1 –** if source includes an actual date of an event	*approxdate* is missing, but *iday and imonth* have values	**1 –** exact date of event is known
2	**2** – if source reports a specific week or the weekend	after 2011: if both *iday* and *approxdate* have a value (note that the full range of days in *approxdate* should be less than a week)	**2 –** the date of the event is known only within a 2-6 day range **3 –** only the week of the event is known
3	**3** – if source reports a particular month without reference to a particular date	1970–2011: if *iday* equals 0 and *imonth* has a value (*imonth*>0) after 2011: if both *imonth* and *approxdate* have a value, but *iday* is 0 or missing.	**4 –** the date of the event is known only within an 8-30 day range or only the month when the event has taken place is known
4	n/a	1970–2011: if *imonth* equals 0.	**5 –** the date of the event is known only within a range longer than one month but not more than one calendar year

B2. LOCATION variables (4 Variables)

Code on following three variables from sources (same variable names all three sources):

10. *latitude*
11. *longitude*
12. *country*

13. *geocoding precision*

Following the temporal precision approach (above), Table 2 illustrates geocoding precision rules for GReVD based on precision variables coded from primary sources, but with some generalization between data sources.

Table 2 Description of the coding values in GReVD and the geocoding precision scale across three sources.

GReVD's coding values	**ACLED** (what *geo_precision* values mean)	**GTD** (what *specificity* values mean)	**UCDP** (what *where_prec* values mean)

1	**1-** the highest precision level	**1-** event occurred in city/village/town and lat/long is for that location	**1-** exact location of the event known and coded.
2	**2-** if the source reports "part of the region" or 'near the town or a city' **3-** if only a larger region is mentioned[18] (code on cases where *admin3* has a value)	**2-** event occurred in city/village/town and no lat/long could be found, so coordinates are for centroid of smallest subnational administrative region identified	**2-** event occurred within at maximum a ca. 25 km radius around a known point. The coded point is the known point.
3	**3-** for cases where *geo_precision* equals 3, if *admin3* is missing, amd the 2nd order administrative unit is known	**3-** event did not occur in city/village/town, so coordinates are for centroid of smallest subnational administrative region identified	**3-** only the second order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid).
4	**3-** for cases where *geo_precision* equals 3, if *admin2* is missing, and the 1st order administrative unit is known	**4-** no 2nd order or smaller region could be identified, so coordinates are for center of 1st order administrative region	**4-** only the first order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid).
5	**3-** for cases where *geo_precision* equals 3, if *admin1* is missing, if no administrative unit is known	**5-** no 1st order administrative region could be identified for the location of the attack, so latitude and longitude are unknown	**5-** the only spatial reference for the event is neither a known point nor a known formal administrative division, but rather a linear feature (e.g. a long river, a border, a longer road or the line connecting two locations further afield than 25 km) or a fuzzy polygon without defined borders (informal regions, large radiuses etc.). A representation point is

[18] According to ACLED Codebook (2019), if a larger region is mentioned, the closest natural location noted in reporting (like “border area”, “forest” or “sea”, among others) is chosen to represent the region – or a provincial capital is used if no other information at all is available – and is noted with precision level “3”. Admittedly, “a larger region” could mean any level of administrative division. Therefore, coding on ACLED’s *admin1, admin2* and *admin3* variables we break down those cases where ACLED’s *geo_precision* equals ‘3’ and expand GReVD’s coding values to ‘2’ through ‘5’, as shown in the table *2.*

			chosen for the feature and employed.
6	n/a	n/a	**6-** only the country where the event took place in is known.
7	n/a	n/a	**7-** event in international waters or airspace.

B3. TAGs variable (1 variable)

14. tags

description: create a “holding” variable in the Registry combining all information on actors, events, means and motives recorded in sources into one variable called *tags* separating by semicolons. A number of variables are captured in *tags* from each of the three sources depending on how sources define and code actors and types of events. This is a text variable.

Rule for ACLED: code on *actor1, actor2, assoc_actor_1, assoc_actor_2, event_type, sub_event_type.*

Rule for GTD: code on *gname, gname2, gname3, targtype1_txt, targtype2_txt, targtype3_txt, attacktype1_txt, attacktype2_txt, attacktype3_txt, suicide.*

Rule for UCDP: code on *side_a, side_b, type_of_violence*.

Below is an example taken from this coding process, drawing on ACLED, GTD and UCDP, to show what a new variable *tags* would look like.

grevd_id	tags
1001000004745705000001	Military Forces of Iraq (2014-2018) Popular Mobilization Forces; Civilians (Iraq); Police Forces of Iraq (2014-2018); Explosions/Remote violence; Shelling/artillery/missile attack
1001000004745705000002	Military Forces of Iraq (2014-2018) Popular Mobilization Forces; Civilians (Iraq); Police Forces of Iraq (2014-2018); Explosions/Remote violence; Shelling/artillery/missile attack
1001000004745705000003	Military Forces of Iraq (2014-2018) Popular Mobilization Forces; Civilians (Iraq); Police Forces of Iraq (2014-2018); Explosions/Remote violence; Shelling/artillery/missile attack
1002201612280045000001	Allied Democratic Forces (ADF); Private Citizens & Property; Armed Assault.
1003000000233552000001	Government of Afghanistan; Taleban; state-based conflict
1003000000233552000002	Government of Afghanistan; Taleban; state-based conflict

B4. SOURCE variables (6 variables)

description: for each entry, code up to three available source materials from primary source for fatality and the dates when source materials were published. Define three source variables *source1, source2* and *source3*. These are string variables. Define three corresponding date variables for each source citation *source1_date, source2_date and source3_date.* It is not expected that each entry will have three sources, but each entry/row will have at least one source available.

15. source1

16. *source2*
17. *source3*

ACLED instructions: code on *source* and *notes*[19] using semicolon as a delimiter*.*

GTD instructions: code on *scite1; scite2; scite3, respectively.*

UCDP instructions: code on *source_article.* All available sources are coded in one variable separating the sources by semicolon. Attempt to extract all sources in order to code them individually. Use the semicolon character as a delimiter.

18. *source1_date*
19. *source2_date*
20. *source3_date*

ACLED instructions: ACLED does not record the date when the source material was published in a separate variable.[20]

GTD instructions: code on *scite1, scite2, scite3* to extract the source dates, respectively.

UCDP instructions: code on *source_date* for events after 2013. Extract the respective dates using semicolon as a delimiter. For events before 2013 attempt to extract the dates from *source_article.*

Codebook references:

ACLED, (2019). "Armed Conflict Location & Event Data Project (ACLED) Codebook."

GTD, (October 2019). "Global Terrorism Database (GTD) Codebook: Inclusion Criteria and Variables"

Högbladh Stina, 2019, "UCDP GED Codebook version 19.1", Department of Peace and Conflict Research, Uppsala University

[19] "If more than two sources are used, the most thorough report is cited, or all are noted in alphabetical order in the 'SOURCE' column" See ACLED Codebook (2019). CHECK with ACLED on extracting sources from *source*.

[20] However, CHECK if possible to extract the date from *notes.*