# UCDP Georeferenced Event Dataset Codebook
# Version 5.0

This version authored by:

**Mihai Croicu,**
**Ralph Sundberg, Ph. D.**

**When citing this dataset, please always cite:**

*The official data presentation article*
Sundberg, Ralph, and Erik Melander, 2013, "Introducing the UCDP Georeferenced Event Dataset", *Journal of Peace Research*, vol.50, no.4, 523-532

*…and the codebook*
Croicu, Mihai and Ralph Sundberg, 2016, "UCDP GED Codebook version 5.0", Department of Peace and Conflict Research, Uppsala University

*The official abbreviation of this dataset is* ***UCDP GED.***
*The official full name of this dataset is* ***UCDP Georeferenced Event Dataset.***
*The current version of the dataset is* ***5.0***

*The current dataset maintainer is* **Mihai Croicu**
*For any questions or comments contact*: **mihai.croicu@pcr.uu.se**

*Data extracted from UCDP systems on 15 October 2016*

# Quick Start Guide:

## What the UCDP GED version 5.0 dataset contains:

The basic unit of analysis for the UCDP GED dataset is the "event", an individual incident (phenomenon) of lethal violence occurring at a given time and place.

More specifically we define an event as: "*An incident where armed force was by an **organised actor** against **another organized actor, or against civilians**, **resulting in at least 1 direct death** at a **specific location** and a **specific date**"*.

The dataset contains **128 264** events. It is a **global dataset** that covers the **entirety of the Globe (excluding Syria)** between **1989-01-01** and **2015-12-31.**

The **maximum (best) spatial resolution** of the dataset is the individual village or town. The dataset is fully geocoded.
The **maximum (best) temporal resolution** of the dataset is the day.

Only events linkable to a UCDP/PRIO Armed Conflict, a UCDP Non-State Conflict or a UCDP One-Sided Violence instance are included. Events are included for the entire period, i.e. both for the years when such conflicts were active and for the years when such conflicts where not active.

The UCDP GED 5.0 is 100% compatible with UCDP GED 4.1, 4.0, 3.0 and 2.0. For backwards compatibility with versions 1.0 – 1.9 see the full codebook.

The dataset is available in **CSV, Shapefile, Rdata, Excel** and **SQL (Postgres, MySQL** and **Microsoft SQL Server)**. A beta public API for GED data is described at: **http://ucdp.uu.se/apidocs/**

## Quick overview of the variables included in the UCDP GED version 5.0:

| Variable name | Content | Type |
|---|---|---|
| **id** | A unique numeric ID identifying each event. | integer |
| **relid** | A quick machine parse-able string key (hash) describing the content of each event.<br><br>The key is constructed using the abbreviation of the country name (for instance AFG for Afghanistan), the calendar year, the type of violence, the dyad or actor ID and a counter.<br><br>This variable is also a unique identifier for each event in | string(255) |

| | | |
|---|---|---|
| | the entire dataset | |
| **year** | The year of the event | integer |
| **active_year** | **1**: if the event belongs to an active conflict/dyad/actor-year<br>**0**: otherwise | Integer |
| **type_of_violence** | Type of UCDP conflict:<br>**1**: state-based conflict<br>**2**: non-state conflict<br>**3**: one-sided violence | integer |
| **conflict_dset_id** | UCDP/PRIO ID for state based conflicts as per the UCDP/PRIO Armed Conflict Dataset.<br><br>UCDP conflict ID code for non-state dyad as per the UCDP Non-State Dataset.<br><br>UCDP actor ID code for the one-sided violence actor as per the UCDP One-Sided Dataset.<br><br>**Note that the same conflict_id in GED can represent two separate conflict, as the dyad_id is unique only within each type of violence (i.e. a nonstate dyad can an identical id with a state-based conflict)[1].**<br><br>**Warning: DO NOT USE this variable to aggregate on or to subset/filter on, as ERRONEOUS results will result. For desired functionality please USE dyad_new_id below. This variable should only be used for merging with data from the UCDP Dyadic, UCDP Non-State and UCDP One-Sided Dataset.**<br><br>Further note that this variable will be deprecated in the future from all UCDP datasets and replaced with the conflict_new_id below. | string(255) |
| **conflict_new_id** | A unique dyad id code for each individual conflict in the dataset.<br><br>**PLEASE DO USE this variable to aggregate, subset or filter by conflict within GED.**<br><br>Further note that this variable will be the main identifier in all future UCDP releases. | integer |
| **conflict_name** | Name of the UCDP conflict to which the event belongs. For non-state conflicts and one-sided violence this is the same as the dyad name. | string(9999) |
| **dyad_dset_id** | UCDP dyad id code for state based dyad as per the UCDP Dyadic Dataset. | string(255) |

[1] For example, dyad id 433 identifies three separate dyads: Government of Sudan – SLM/A (state-based), Government of Senegal – civilians (one-sided) and Dioula – Guere (non-state)

| | | |
|---|---|---|
| | UCDP conflict ID code for non-state dyad as per the UCDP Non-State Dataset.<br><br>UCDP actor ID code for the one-sided violence actor as per the UCDP One-Sided Dataset.<br><br>**Note that the same dyad_id in GED can represent two separate dyads, as the dyad_id is unique only within each type of violence (i.e. a nonstate dyad can an identical id with a state-based dyad).**<br><br>**Warning: DO NOT USE this variable to aggregate on or to subset/filter on, as ERRONEOUS results will result. For desired functionality please USE dyad_new_id below. This variable should only be used for merging with data from the UCDP Dyadic, UCDP Non-State and UCDP One-Sided Dataset.**<br><br>Further note that this variable will be deprecated in the future from and replaced with the dyad_new_id below. | |
| **dyad_new_id** | A unique dyad id code for each individual dyad in the dataset.<br><br>**PLEASE DO USE this variable to aggregate, subset or filter by dyad within GED.** | integer |
| **dyad_name** | Name of the conflict dyad creating the event.<br><br>A **dyad is the pair of two actors engaged in violence** (in the case of one-sided violence, the perpetrator of violence and civilians).<br><br>The two sides are separated by an ASCII dash (e.g. Government of Russia - Caucasus Emirate, Taleban – civilians). | string(9999) |
| **side_a_dset_id** | The unique ID of side A as in the current version of the UCDP Actor Dataset | string(255) |
| **side_a_new_id** | A unique ID of side A. This will be used in the future as the main actor identifier. | integer |
| **side_a** | The name of Side A in the dyad. In state-based conflicts always a government. In one-sided violence always the perpetrating party. | string(9999) |
| **side_b_dset_id** | The unique ID of side B as in the current version of the UCDP Actor Dataset | string(255) |
| **side_b_new_id** | A unique ID of side B. This will be used in the future as the main actor identifier. | integer |
| **side_b** | The name of Side B in the dyad. In state-based always the rebel movement or rivalling government. In one-sided violence always “civilians”. | string(9999) |
| **number_of_sources** | Number of total sources containing information for an event that were consulted. | integer |

| | | |
|---|---|---|
| | Note that this variable is only available for data collected for 2013 and 2014, and for recently revised events. For older data, -1. Note that -1 does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field. | |
| **source_article** | References to the names, dates and titles of the source material from which information on the event is gathered.<br><br>**A reference to at least one source material is available for ALL EVENTS.**<br><br>This variable is highly streamlined for information collected in 2013 and 2014, and less so for older data. For such older data, abbreviations such are used. The most frequent are:<br><br>R: Reuters News,<br>BBC: BBC Monitoring<br>AP: Associated Press Newswires<br>AFP: Agence France Presse,<br>X: Xinhua<br>DOW: Dow Jones Wires | text |
| **source_office** | The name of the organizations publishing the source materials.<br><br>Note that this variable is only available for data collected for 2013 and 2014, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for every event. | text |
| **source_date** | The dates the source materials were published.<br><br>Note that this variable is only available for data collected for 2013 and 2014, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for every event. | text |
| **source_headline** | The titles of the source materials.<br><br>Note that this variable is only available for data collected for 2013 and 2014, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for | text |

| | | |
|---|---|---|
| | every event. | |
| **source_original** | The name or type of person or organization from which the information about the event originates in the original report.<br><br>e.g. “police”, “Lt. Col. Johnson”, ”eyewitnesses”, “rebel spokesman”. | string(9999) |
| **where_prec** | The precision with which the coordinates and location assigned to the event reflects the location of the actual event.<br><br>**1**: exact location of the event known and coded.<br>**2**: event occurred within at maximum a ca. 25 km radius around a known point. The coded point is the known point.<br>**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).<br>**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).<br>**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 chosen for the feature and employed.<br>**6**: only the country where the event took place in is known.<br>**7**: event in international waters or airspace. | integer |
| **where_coordinates** | Name of the location to which the event is assigned. Fully standardized and normalized. | string(9999) |
| **where_description** | An extracted snippet of text from the source material describing the location. | text |
| **adm_1** | Name of the first order (largest) administrative division where the event took place | string(9999) |
| **adm_2** | Name of the second order administrative division where the event took place | string(9999) |
| **latitude** | Latitude (in decimal degrees) | numeric(9,6) |
| **longitude** | Longitude (in decimal degrees) | numeric(9,6) |
| **geom_wkt** | An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as *well known text* without SRID. | string(9999) |
| **priogrid_gid** | The PRIO-grid cell id (**gid**) in which the event took place. Compatibility with PRIO-grid (Tollefsen, 2012) is guaranteed for both PRIO-grid **1** and **2**. | integer |

| | | |
|---|---|---|
| | **Warning:** We associate every point to the PRIO-grid that contains it, even if the point is in another country than the one officially assigned to the respective PRIO-grid cell through their majority area rule. It is your responsibility to make sure the covariates for the PRIO-grid cell are correct for each event. Further, for the same reason, DO NOT, under any circumstances, first clip out (subset) PRIO-grid by country before merging with UCDP GED as data loss will certainly occur. Refer to your copy of the PRIO-grid for further details on PRIO-grid's majority assignment rule (p.3). | |
| **country** | Name of the country in which the event takes place. | string(999) |
| **region** | Region where the event took place. One of following: {Africa, Americas, Asia, Europe, Middle East} | string(999) |
| **event_clarity** | 1 (high) for events where the reporting allows the coder to identify the event in full. That is, events where the individual happening is described by the original source in a sufficiently detailed way as to identify individual incidents, i.e. separate activities of fighting in a single location:<br><br>Example of such reporting: *"2 people where killed in Banda Aceh town on the 9th of December in fighting between the government and GAM when a car exploded in a main market."*<br><br>2 (lower) for events where an aggregation of information was already made by the source material that is impossible to undo in the coding process. Such events are described by the original source only as aggregates (totals) of multiple separate activities of fighting spanning over a longer period than a single, clearly defined day.<br><br>Examples of such reporting: "The Ukrainian government informs that 29 people have died in the past six days in a number of clashes with the separatists along the line of conflict". | integer |
| **date_prec** | How precise the information is about the date of an event.<br>**1**: exact date of event is known;<br>**2**: the date of the event is known only within a 2-6 day range.<br>**3**: only the week of the event is known<br>**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<br>**5**: the date of the event is known only within a range longer than one month but not more than one calendar year. | integer |
| **date_start** | The earliest possible date when the event has taken place. | Date YYYY-MM- |

| | | |
|---|---|---|
| | | DD |
| **date_end** | The last possible date when the event has taken place. | Date YYYY-MM-DD |
| **deaths_a** | The best estimate of deaths sustained by side a.<br><br>Always 0 for one-sided violence events. | integer |
| **deaths_b** | The best estimate of deaths sustained by side b.<br><br>Always 0 for one-sided violence events. | integer |
| **deaths_civilians** | The best estimate of dead civilians in the event.<br><br>For non-state or state-based events, this is the number of collateral damage resulting in fighting between side a and side b. For one-sided violence, it is the number of civilians killed by side a. | integer |
| **deaths_unknown** | The best estimate of deaths of persons of unknown status. | integer |
| **best_est** | The best (most likely) estimate of total fatalities resulting from an event.<br>It is always the sum of **deaths_a, deaths_b, deaths_civilians and deaths_unknown**. | integer |
| **high_est** | The highest reliable estimate of total fatalities | integer |
| **low_est** | The lowest reliable estimate of total fatalities | integer |
| **geom / geometry** | An Open Geospatial Consortium / ESRI binary representation of each individual point. Contains the SRID (4326) where supported.<br><br>Due to the binary nature of this variable, this variable is contained only in the formats that support it. | geometry (Point,4326) |
| **isocc** | The ISO 3-letter code of the country where the event takes place. | string(3) |
| **gwno** | The Gleditsch and Ward numeric identifier of the country where the event takes place | integer |
| **gwab** | The Gleditsch and Ward alphabetic identifier of the country where the event takes place | string(3) |

This ends the quick-start guide. For more detailed information, please refer to the detailed dataset description in the next chapters of the codebook.
Remainder of the page left empty for your notes.

# 1 : Statement of Purpose

The purpose of this project is to provide the academic community with the most comprehensive structured event data available on organised violence in the post-1989 world, so as to answer the call for geographically and temporally disaggregated data.

Whereas the ambition is to provide a dataset with both theoretical and practical relevance for researchers in a broad range of scholarly traditions, mainly pragmatic and practical decisions guide the construction of the dataset. This allows for effective coding procedures as well as disaggregated and flexible data without predetermined biases for certain research purposes. The geo-referenced event data may thus be used for purposes ranging from wanting to illustrate conflict behaviour geographically, using geographic information systems software, to studying causal pathways by applying a variety of methods for statistical analysis.

Whilst retaining the ambition to provide a dataset open for a broad variety of research purposes, the focus of the dataset on conflict dynamics and the effects of armed violence, in the form of deaths, still sets the parameters for users. This means that the UCDP GED is in effect primarily directed toward, and will most probably be useful to, quantitative and comparative researchers interested in the fatal outcomes of violent conflict behaviour at the level below the state.

**Thus, the dataset is constructed in such a way as maximize the comparability and consistency across time and space, and provide a globally consistent image of the phenomenon of organized violence.**

**In effect, the goal of UCDP GED is not to present the most complete and accurate image of a certain conflict at a certain point in time, but rather be a tool for the global understanding of subnational conflict patterns and trends.**

**Version 5 is the first full global event dataset, and consists of the data for the whole World between 1989 and 2015, including cases such as Iraq, Colombia, Pakistan, Afghanistan and India.** Data for Syria is not included in the current version – as data collection processes and procedures at this level of disaggregation did not yield a product releasable at this time with the same level of consistency and clarity as other GED data.

# 2 : Definitions

We define an event as:

*An incident where **armed force** was used by an **organised actor** against **another organized actor, or against civilians**, **resulting in at least 1 direct death** at a **specific location** and a **specific date**".*

These are the specific elements of the definition:

1. ***Armed force***: use of *arms* in order to promote the parties' general position in the conflict, resulting in deaths.

- arms: any material means e.g. manufactured weapons but also sticks, stones, fire, water etc.

2. ***Organized actor***: a government of an independent state, a formally organized group or an informally organized group according to UCDP criteria:

a. ***Government of an independent state***: The party controlling the capital of a state.

b. ***Formally organized group***: Any non-governmental group of people having announced a name for their group and using armed force against a government (state-based), another similarly formalized group (non-state conflict) or unorganized civilians (one-sided violence). The focus is on armed conflict involving consciously conducted and planned political campaigns rather than spontaneous violence.

*c.* ***Informally organized groups***: Any group without an announced name, but which uses armed force against another similarly organized group (non-state conflict), where the violent activity indicates a clear pattern of violent incidents that are connected and in which both groups use armed force against the other

3. ***direct death*** : a death relating to either combat between warring parties or violence against civilians.

UCDP GED provides three estimates for deaths for each event, thus creating an uncertainty interval:

- a low estimate, containing the most conservative estimate of deaths that is identified in the source material;

- a best estimate, containing the most reliable estimate of deaths identified in the source material;

- a high estimate, containing the highest reliable estimate of deaths identified in the source material. Note that UCDP attempts to distinguish and not include unreasonable claims in the high estimate of fatalities, and tends to be highly conservative when counting fatalities[2].

In order for an event to exist, at least one dead needs to be registered in the high, best or low estimate.

4. ***Specific location***: a name and one pair of latitude and longitude coordinates that relate to the geographical information specified in the source material.

5. ***Specific date***: a specified time period during which armed interactions cause at least 1 fatality. The normal temporal unit to which an event can be related is a 24-hour day starting at midnight.

- In some cases it is impossible, based on the source material, to reduce the specific date to a single day as reporting only refers to wider time spans (multiple

[2] For a more elaborate discussion on aspects concerning point 1-4, please refer to UCDP Codebooks for State-Based Armed Conflicts, Non-state Conflicts and One-Sided Violence.

days) or information on the exact day is not clear. For these events, a wider time span is provided through the use of the *date_start, date_end and date_prec* variables.

For further UCDP definitions please refer to the "Definitions" section of UCDPs webpage available at http://www.pcr.uu.se/research/ucdp/definitions/

Note that this definition is fully compatible with the one used for GED 1.0-4.1.

# 3. Comparison with other event ontologies:

The **UCDP GED is an "incident dataset",** sharing a highly similar conceptualization of "events" with datasets such as the Global Terrorism Dataset (START, 2013), ACLED (Raleigh, 2009) or SCAD (Salehyan, 2012) in the sense that each entry represents an **incident,** a real-life series of actions circumscribed to a certain typology and resulting in a certain outcome or set of outcomes. In the case of UCDP GED this typology is: **fighting resulting in the death of at least one person**.

This differs markedly from the conceptualization of events in most "machine" datasets and coding systems such as PHOENIX/EL:DIABLO (OEDA, 2014), CAMEO/TABARI (Gerner et. al, 2014) , ICEWS/JABARI (Boschee et. al., 2015) or KEDS (Schrodt, 2006). In these datasets, an event represents an action between two actors (e.g. taunt, attack, fight, retreat, mediate etc.). Since typically multiple actions lead to an outcome (an incident) a UCDP GED event is equivalent to a collection or aggregation of multiple events in such "machine" datasets. Further, as compared to these datasets, UCDP GED is more geared towards extracting information relating to the incident (such as fatality figures) rather than binning actions into sets of categories.

Further, UCDP GED differs markedly to event data collection efforts where the event is defined as the individual source report, such as for example, MMAD (Rød and Weidmann, 2014), since UCDP GED collates information from all sources referring to the same information and interprets all of them in order to extract information and define an event. Thus, UCDP GED would be a filtered version of such a dataset, with filtering performed according to certain interpretation rules, uncertainty management and definitions.

Of course, this comparison only describes the conceptual differences between our conceptualization of "events" and other conceptualization of "events" for users either more familiar with other datasets or users wanting to use UCDP GED together with other event-type data. As such, no other event dataset replicates or is replicated by UCDP GED as UCDP GED collects information on a specific, unique type of social behaviour (as defined above).

# 4. Sources and data collection process:

The UCDP GED is manually curated and compiled, with automatic assistance in data retrieval, filtering, data storage and manipulation, as well as data validation.

The original reporting underlying UCDP GED are collected from three sets of sources:

1. **global newswire reporting**
2. **global monitoring and translation of local news performed by the BBC**
3. **secondary sources such as local media, NGO and IGO reports, field reports, books etc.**

with a slight majority (approximately 60% of the dataset) of all events being based on global newswires reporting.

The process is done in a "two-pass" system, first by consulting newswire sources for the entire globe then by consulting local/specialized sources based on information obtained from the first pass.

Further, the UCDP GED is based on the same underlying data that all the other UCDP datasets are based on, i.e. UCDP GED is not built FROM e.g. the UCDP Dyadic Dataset, but rather both the UCDP Dyadic Dataset and GED are built from the same data.

## 4.1 First pass: Global Newswire and BBC Monitoring

**Global newswire** reporting as well as **BBC Monitoring data** are sourced from the **Dow Jones Factiva** aggregator, using the following general search-string:

```
kill* or die* or injur* or dead or death* or wounded or massacre*
```

This search is done globally, with "head and lead" and "intelligent indexing" being used for further filtering in those cases where this is feasible.

In terms of sources, UCDP uses Reuters News, Agence France Presse (in English), Associated Press, Xinhua (in English) as well as BBC Monitoring. Note that for some of the years and geographic areas, reporting for some of the sources is extremely limited due to Factiva's non-inclusion of the whole corpus.

Similarly, media reporting is not consistent across time or space for any of the above-mentioned organizations. Changing managerial focuses, different organizational structures (such as field office locations), as well as different resource distributions and allocations (such as, for example, the restructuring of BBC Monitoring in the early 2010s) make media reporting quality and quantity vastly different over various periods and over different areas. Furthermore, the ways in which conflicts are reported set the parameters for the preciseness of the data. In some countries and some phases of conflicts, the event data is based on either detailed daily reports or more summary-like reports covering larger areas.

Given this vast inconsistency in the source material generation process, we do not aim to be "source-consistent" - i.e. we do not aim to use the exact same set of sources to generate the entire dataset, since this would not provide anything else than a very poor convenience sample. The non-randomness in such a sample would be driven by news-agencies' procedures.

Instead, we drive our coding procedures with the set goal of obtaining a population dataset at the aggregate level (a complete yearly list) of all UCDP conflicts. This allows for consistency to be passed down to the event level and thus increases the quality and reliability of the sample presented in GED[3].

## 4.2 Second pass: Local and specialized news-sources

Thus, for many conflicts and periods we add: **secondary sources** such as:

- **local monitoring of various local media** (e.g. Press Trust of India for India, or EFE news agency for Latin America or Radio Okapi for DR. Congo),
- **local monitoring and research organizations** (such as SATP for India and Pakistan),
- **global NGO reports** (such as those coming from Human Rights Watch or Amnesty International),
- **UN, EU, AU and other IGO reports**,
- **governmental publications** (where considered reliable, such as those from Truth and Reconciliation Commissions),
- **research articles or books etc**.

The choice of whether to include secondary sources is made by project leaders and UCDP coders together at the conclusion of the first pass. The goal of this inclusion is to further specify and identify conflicts (at the aggregate country-year level) in places where detail is seen as insufficient. Thus, by attempting to equalize the level of detail identified across the board we make the sample more normalized for substantive analysis.

Coders are experts in the coding process. Unlike any other data collection project in the field, all coders are full-time long-term employees of UCDP, typically following conflicts and countries for long periods, and attaining in many cases specialist status in certain geographical areas. Further, UCDP consults external area specialists as to best select sources appropriate for this second pass.

## 4.3 Further considerations for data validity and reliability

In general, the codebook and its appendices aim to contribute to improve, as much as possible, the reliability of the data, by presenting clear and consequent definitions as well as transparent coding procedures and rules.

[3] While we do claim that our approach will provide a better sample for substantive analysis in most cases, we recognize that some analyses will require a source-consistent dataset. If you do need to subset the data for source consistency (with the above-mentioned caveats), you can subset the dataset to solely use one or all news-sources using the ***source_article*** variable. You may find such subsets valuable for things like analyses on media biases or, in conjunction with the full dataset, capture-recapture tests.

The constructed precision codes for time, geography and event clarity, however detailed and elaborated, may allow for differing interpretations and understandings. Though coding rules and precision codes have been extensively discussed with researchers and tested in a pilot phase of the project during the summer of 2009, the process of constructing the geo-referenced event dataset is based on several procedures that may not always correspond to the reality of the events. For example when constructing the dataset, the UCDP coders have, for pragmatic reasons, worked from the assumption that all events referring to the same start and end dates and 1 location represent an event clarity of 1. However, due to changing coding rules over a long period of time for the annual UCDP data, some of the dates as well as the included information are not as precise as others. This is especially true for the years 2002 and 2003 during which the UCDP experienced major structural rearrangements and improvements.

Further, differences in reporting may affect not only how much of the real world population of events is coded (see above), but also what detail level can be extracted. As such, for some countries, precise locations might be uncommon in reports on armed violence. There might even be a preference towards reporting violent activities on the first-order administrative level or less, which decreases the geographical precision in large.

In relation to this, the coders of the GED are experts on the coding procedure, yet seldom on the geographical dimensions of each conflict. This opens up for an error marginal where unclear location phrases such as “area” or “zone” can be misinterpreted. To address this challenge, the UCDP begins with studying the geographical and administrative structures for each new country to code.

## 4.4. Quality assurance

The dataset includes an extensive series of procedures to assure the quality and reliability of the data.

First, a two-stage coding procedure is employed for each event, with at least two separate coders being in charge of coding and revising each individual event for the finalized product. This two-stage procedure is conducted at different times (for most events at least one year apart, but in other cases as much as 10 years apart), and uses separate sets of procedures, so as to insure that coders do not influence each-other and to insure inter-coder reliability.

A large number of routines are set in place at this stage for quality control, each coder being given a fixed, comprehensive set of protocols to be followed that ensure the consistent treatment of, amongst other, dyad names, dyad IDs, precision scores, geocoding locations, streamlining of names, integrity of fatality estimates etc. The exact coding procedure is highly formalized, with all the steps of the process being given to coders, together with algorithms to insure the correctness of all the decisions that coders have to take. Similarly, the assignation of each event to a type of violence and to a dyad is done by at least two coders following a similarly strict set of routines.

Identified inconsistencies are resolved through regular, frequent meetings (at least once a week) where all the coders and project managers take part.

Third, over 50 automatic tests are applied to the data, followed by a series of manual checks by a project leader. Algorithms verifying that locations are properly geo-coded

(through the usage of density analyses and interpolation techniques), that ADM1s and ADM2s are properly identified and linked to events, that death estimates are properly used, that IDs are properly used and consistent with the aggregated datasets are done etc. Some of these tests are done at point of input, GED using a state-of-the art, custom built data-management, data-input and data-storage facility, others at release time.

Visualizations of the data are provided to coders and project managers using a Google Maps API derived solution.

The automated routines do not make any modifications to the datasets, requiring a human coder to make all the changes for an added level of security. The automated tests are re-run for as many times as required, until the data is deemed as acceptable for release by a project manager.

**** INSERT SOME TEXT ABOUT External Assessments and Validations ****

# 5. Data inclusion:

The event dataset has a **dyad and actor focus**, tracing the events of all UCDP conflict dyads[4] for both active years (years that have crossed the 25 battle related deaths threshold) and non-active years (the remainder).

Thus, if a dyad crossed the 25-deaths threshold in a single year, but did generate some events in either previous or subsequent years, all events belonging to the dyad are included, including those in years where the threshold was not crossed[5].

The dataset includes all three types of UCDP organised violence: state-based conflict, non-state conflict and one-sided violence. All three categories of the UCDP annual data are mutually exclusive and coded events will therefore also be exclusive and non-overlapping. The data series start in 1989 and events before this calendar year are not included.

All the inclusion criteria are identical to UCDP GED version 1.5 – 4.1.

[4] A dyad consists of two conflicting primary parties or party killing unarmed civilians. In state-based armed conflicts, at least one of the primary parties must be the government of a state.
A state-based conflict can include more than one dyad, if multiple groups oppose the government over the same incompatibility; non-state conflicts and one-sided violence instances are always equivalent to a dyad.
Further, in non-state armed conflicts, a dyad can only consist of formally versus formally organized groups or informally versus informally organized groups. A formally organized group can not be fighting an informally organized group to keep non-state conflicts and one-sided violence as independent categories.

[5] E.g. State-based dyad 431 (Government of Uganda – UNRF II) crosses the 25 battle-related deaths threshold only in 1997. However, this dyad had some events, but did not cross the 25 battle-related deaths in 1996 and 1998. In versions 1.0 and 1.1 only those events belonging to the dyad in 1997 were included. In this version all the events belonging to the dyad (including those in 1996 and 1998) were included.

For a tabular view of how the UCDP GED dataset relates to other UCDP products please refer to table 2 below. Note that the release of UCDP GED is not synchronized perfectly with the above datasets, thus data discrepancies may appear due to data revisions.

**Table 2.** UCDP data corresponding to the the GED

| Dyad Type | Period | Actor Inclusion | Event Inclusion | Reference |
|---|---|---|---|---|
| *State-Based* | 1989-2015 | All dyads that cross the 25 battle-related deaths threshold in at least one year of the period and have a stated goal of incompatibility.<br><br>The whole activity of these dyads is included over the entire 1989-2015 period (including those dyad-years when the 25 battle related deaths threshold was not exceeded). | All events leading to at least one death between:<br><br>1989-01-01 to 2015-12-31 | UCDP/PRIO Armed Conflict Dataset Codebook Version 4-2016 |
| *Non-State* | 1989-2015 | All dyads that cross the 25 battle-related deaths threshold in at least one year of the period.<br><br>The whole activity of these dyads is included over the entire 1989-2015 period (including those dyad-years when the 25 deaths threshold was not exceeded).. | All events leading to at least one death between:<br><br>1989-01-01 to 2015-12-31 | UCDP Non-State Conflict Codebook Version 2.5-2016 |
| *One-Sided* | 1989-2015 | All dyads that have ever crossed the 25 deaths threshold during this period.<br><br>The whole activity of these dyads is included over the entire 1989-2015 period (including those actor-years when the 25 deaths threshold was not exceeded). | All events leading to at least one death between:<br><br>1989-01-01 to 2015-12-31 | UCDP One-Sided Violence Codebook, Version 1.4- 2016 |

# 6. Dataset content:

The dataset contents can be divided into 7 categories: **event identifiers; actors and dyads; sources**; **geography**; **time**; **clarity**; **fatality figures.**

Note that the current version of the dataset contains different field identifiers (variable names) and field positions than GED version 1. You will have to adapt your scripts

developed for UCDP GED 1 to work with UCDP GED 2-5. Further note that some variables have changed, and others have been added.

All variables present in **UCDP GED 5** have been present in **UCDP GED 2, 3, 4 and 4.1.**

Variables new in GED 2/3/4 as compared to GED 1 are **underlined in red.**
Variables changed in GED 2/3/4 as compared to GED 1 are **underlined in black.**
Variables unchanged throughout GED 1/2/3 are **in simple bold.**

## 6.1 Event identifiers

This section provides unique identifiers for every event (row/entry) in the dataset. All variables in this section can be used as a unique key for the dataset.

| | | |
|---|---|---|
| **id** | **A persistent unique numeric ID identifying each event.**<br><br>The same id in versions 1.9, 2, 3, 4, 4.1 and 4.5 identifies the same event (incident). This allows changes between versions to be traced at event level. | integer |
| **relid** | A quick machine parse-able string key (hash) describing the content of each event.<br><br>The key is constructed using the abbreviation of the country name (for instance AFG for Afghanistan), the calendar year, the type of violence, the dyad or actor ID and a counter.<br><br>The format of the key is CCC-YYYY-T-DDDDD-SSSS.<br><br>CCC = Gleditsch and Ward 3-letter textual country code for the first country of activity of side A (e.g. AFG for Afghanistan).<br><br>YYYY = Calendar Year of the event.<br><br>T = Type of violence (1 for state-based violence, 2 for non-state violence, 3 for one-sided violence).<br><br>DDDD = Dyad_id of the event.<br><br>SSSSS = A counter. | string(255) |

## 6.2 Actors and dyads

This section provides variables that allow for linkages between the UCDP GED and all other UCDP datasets.

This section also provides with variables to allow you to aggregate/filter/extract data on conflict, dyad or actor.

| | | |
|---|---|---|
| **active_year** | 1**: if the event belongs to an active conflict/dyad/actor-year**<br>0**: otherwise** | **Integer** |
| **type_of_violence** | Type of UCDP conflict:<br>**1**: state-based conflict<br>**2**: non-state conflict<br>**3**: one-sided violence | integer |
| **conflict_dset_id** | UCDP/PRIO ID for state based conflicts as per the UCDP/PRIO Armed Conflict Dataset.<br><br>UCDP conflict ID code for non-state dyad as per the UCDP Non-State Dataset.<br><br>UCDP actor ID code for the one-sided violence actor as per the UCDP One-Sided Dataset.<br><br>**Note that the same conflict_id in GED can represent two separate conflict, as the dyad_id is unique only within each type of violence (i.e. a nonstate dyad can an identical id with a state-based conflict)[6].**<br><br>**Warning: DO NOT USE this variable to aggregate on or to subset/filter on, as ERRONEOUS results will result. For desired functionality please USE dyad_new_id below. This variable should only be used for merging with data from the UCDP Dyadic, UCDP Non-State and UCDP One-Sided Dataset.**<br><br>Further note that this variable will be deprecated in the future from all UCDP datasets and replaced with the conflict_new_id below. | string(255) |
| conflict_new_id | A unique dyad id code for each individual conflict in the dataset.<br><br>**PLEASE DO USE this variable to aggregate, subset or filter by conflict within GED.** | integer |
| **conflict_name** | Name of the UCDP conflict to which the event belongs. For non-state conflicts and one-sided violence this is the same as the dyad name. | string(9999) |
| **dyad_dset_id** | UCDP dyad id code for state based dyad as per the UCDP Dyadic Dataset.<br><br>UCDP conflict ID code for non-state dyad as per the UCDP Non-State Dataset. | string(255) |

[6] For example, dyad id 433 identifies three separate dyads: Government of Sudan – SLM/A (state-based), Government of Senegal – civilians (one-sided) and Dioula – Guere (non-state)

| | | |
|---|---|---|
| | UCDP actor ID code for the one-sided violence actor as per the UCDP One-Sided Dataset.<br><br>**Note that the same dyad_id in GED can represent two separate dyads, as the dyad_id is unique only within each type of violence (i.e. a nonstate dyad can an identical id with a state-based dyad).**<br><br>**Warning: DO NOT USE this variable to aggregate on or to subset/filter on, as ERRONEOUS results will result. For desired functionality please USE dyad_new_id below. This variable should only be used for merging with data from the UCDP Dyadic, UCDP Non-State and UCDP One-Sided Dataset.** | |
| dyad_new_id | A unique dyad id code for each individual dyad in the dataset.<br><br>**PLEASE DO USE this variable to aggregate, subset or filter by dyad within GED.** | integer |
| **dyad_name** | Name of the conflict dyad creating the event.<br><br>A **dyad is the pair of two actors engaged in violence** (in the case of one-sided violence, the perpetrator of violence and civilians).<br><br>The two sides are separated by an ASCII dash (e.g. Government of Russia - Caucasus Emirate, Taleban – civilians). | string(9999) |
| **side_a_dset_id** | The unique ID of side A as in the current version of the UCDP Actor Dataset | string(255) |
| side_a_new_id | A unique ID of side A. | integer |
| **side_a** | The name of Side A in the dyad. In state-based conflicts always a government. In one-sided violence always the perpetrating party. | string(9999) |
| **side_b_dset_id** | The unique ID of side B as in the current version of the UCDP Actor Dataset | string(255) |
| side_b_new_id | A unique ID of side B. | integer |
| **side_b** | The name of Side B in the dyad. In state-based always the rebel movement or rivalling government. In one-sided violence always “civilians”. | string(9999) |

## 6.3. Description of Sources

This section contains references to the sources underlying each event. See section 4.2 for a description of the data collection processes and source selection process.

While not available in the dataset, UCDP does keep the full text of the underlying reporting. As most of this text is copyrighted to news agencies/publishers, it is impossible to supply such texts together with the dataset. If you need to obtain access to the full text of reports, you will either need to re-download them from Factiva/Lexis Nexis[7].

UCDP does not store the unique identifiers that Factiva, Reuters, AFP etc. assigns to events, as during the decades-long data collection process we observed such identifiers change multiple times, making them useless for tracing source material directly.

| | | |
|---|---|---|
| number_of_sources | **Number of total sources containing information for an event that were consulted.**<br><br>**Note that this variable is only available for data collected for 2013 - 2015, and for recently revised events. For older data, -1. Note that -1 does** NOT **mean information on the source is missing; reference to the source material is** ALWAYS **available in the** source_article **field.** | **integer** |
| **source_article** | References to the names, dates and titles of the source material from which information on the event is gathered.<br><br>**A reference to at least one source material is available for ALL EVENTS.**<br><br>This variable is highly streamlined for information collected in 2013 - 2015, and less so for older data. For such older data, abbreviations such are used. The most frequent are:<br><br>R: Reuters News,<br>BBC: BBC Monitoring<br>AP: Associated Press Newswires<br>AFP: Agence France Presse,<br>X: Xinhua<br>DOW: Dow Jones Wires | text |
| source_office | The name of the organizations publishing the source materials.<br><br>Note that this variable is only available for data collected for 2013 - 2015, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for every event. | text |

[7] For very small samples or original reports or information on individual events, you are welcome to contact us.

| | | |
|---|---|---|
| source_date | The dates the source materials were published.<br><br>Note that this variable is only available for data collected for 2013 - 2015, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for every event. | text |
| source_headline | The titles of the source materials.<br><br>Note that this variable is only available for data collected for 2013 - 2015, and for recently revised events. For older data, the field is empty. Note that an empty field does **NOT** mean information on the source is missing; reference to the source material is **ALWAYS** available in the **source_article** field, for every event. | text |
| **source_original** | The name or type of person or organization from which the information about the event originates in the original report.<br><br>e.g. “police”, “Lt. Col. Johnson”, ”eyewitnesses”, “rebel spokesman”. | string(9999) |

## 6.4. Geography

Data in the UCDP GED is geo-referenced, meaning that each event is connected to a specific location defined by a pair of latitude and longitude coordinates.

**Each event is connected to a single location**. If reporting talks about multiple locations but gives only one aggregated fatality figure is given, then the following procedure is applied**:**

- one separate event is created for each location;
- deaths are split between locations as evenly as possible in order to maintain the fatality figures as integers. The split is performed automatically by the data management system[8].

The coordinates are fixed to the World Geodetic System of 1984 (WGS 84), EPSG SRID 4326. These coordinates are specified in decimal degrees with a precision of 6 decimal figures (e.g. 75.920211). Coordinates (latitude and longitude) used in the GED are based on the most precise location mentioned in the source.

**The lowest level of spatial disaggregation for an urban location is the town, for the rural areas, the village.**

[8] if insufficient deaths exist to create the required number of events (e.g. the reporting speaks of three locations and of only two dead), no split is performed. Instead, the smallest geographical unit encompassing all mentioned locations is used with an appropriate precision score.

Street, neighborhoods, parts of towns are not coded, even when such information is available in the reporting. Thus, a town is always represented by a single pair of latitude and longitude coordinates.

Suburbs, as long as they can be seen as separate urban areas, distinct from the main town, are coded as individual towns. Similarly, airports are always coded as separate entities.

Other features such as "mountains", "peaks" and "forests" are also used to specify geographical location, as long as their size is comparable (same order of magnitude) to those of towns or villages.

**The next lowest levels of spatial disaggregation are the administrative division of the country.**

UCDP uses two levels administrative divisions for every country, the first-order administrative division (referred to as the ADM1) and the second order administrative division (referred to as the ADM2).

In the case of multiple, contested administrative systems (such as in Sri Lanka or Nagorno-Karabakh), UCDP uses the administrative system of the government controlling the capital of the country where the event takes place in.

**The highest level of spatial aggregation for location is the country, defined using the Gleditsch and Ward list**.

Further, all the geocoding is **time-aware,** i.e. locations are coded to the place-names and administrative divisions that were in place at the time the event took place. For example, an event that took place in 1989 in what is today St. Petersburg, Russia, is geocoded as happening in Leningrad, Soviet Union. Thus, changes in administrative structures of countries, as well as changes in borders are visible in UCDP GED.

All text-based information on location is provided in two ways:

- the name of the location as mentioned in the text (`where_description`).
- the name of the location whose coordinates were assigned to the event (`where_coordinates`).

While in most cases the two are similar, in two cases they will differ:

- if the location mentioned is not found during the geocoding process, and a coarser administrative unit is used (ex. if Ngwan village in Shan State is found in the text, but the coder cannot find that exact village, `where_description` will indicate Ngwan village in Shan State whereas `where_coordinates` will indicate Shan State. The latitude and longitude in this case will be those of Shan State).

- if the location is spelled in different ways (e.g. Mumbai and Bombay; in `where_description` the spelling of the article will be used, in `where_coordinates` only one form will be present: Mumbai town).

`where_coordinates` is always streamlined - a latitude/longitude pair will only ever link to one `where_coordinate`. Further in `where_coordinate`, all capitals are

referred to as “cities”, all urban localities other than capitals as “towns” (New York City Town is a correct name in where_coordinate), all rural localities as villages or localities etc.

### 6.4.1. Geo-referencing sources

UCDP does not employ an over-arching source for geocoding, as experience has proven that there is no quality global source for location data, especially for conflict zones and least-developed countries.

As such, UCDP coders employ sources such as global gazetteers (such as the United States National Geospatial Intelligence Agency’s GEOnet Names Server, Maplandia, GeoHack or the Google Geocoding API), local maps provided by governmental authorities, UN agencies (such as UN OCHA) or local NGOs, as well as, on occasion, historical maps such as the US Army Map Service Global Topographic Maps series.

Supervised semi-automatic geocoding is employed in a number of cases (mainly in Europe and the Former Soviet Union), using Google Geocoding API, Yandex and Bing. Strings to be geocoded are always manually extracted, however, and the resulting geocoding is vetted both manually and by automatic procedures.

Extreme care is taken to insure the full consistency, coherence and reliability of the data across the dataset. UCDP maintains both a repository of all the names previously geo-coded, as well as internal automated systems designed to insure that consistency (such as 1:1 matches between place-names and coordinates) is maintained throughout the dataset.

Information used to determine administrative divisions (labelled ADM1 and ADM2) stem from several different sources, commonly from a government’s own website or reference literature that covers administrative divisions globally. The global ISO 3166-2 standard is further used for identifying administrative divisions.

Note that while in most cases ADM1s are the largest administrative divisions in a country, in some cases (such as Russia or Romania) they are not, as the largest administrative division is either solely a statistical reporting unit or simply a legal fiction.

Correspondence regarding geographical coordinates, administrative divisions and any general questions or comments regarding the geographic aspects of the coding should be emailed to the maintainer of the dataset. Also, please report any potential errors in the dataset.

### 6.4.2. Geo-precision and its Values

In order to determine the precision with which specific latitude and longitude coordinates are connected to an event location, the dataset uses a geo-precision variable. Precise coding rules and examples of how the geo-precision values are assigned in the GED can be found in the Appendix.

The geo-precision variable can have seven values:

1 - Event can be related to an exact location, meaning a place name with a specific pair of latitude and longitude coordinates;

2 - Event can be "near", in the "area" of or up to 25 km away from an exact location, meaning a place name with a specific pair of coordinates;

3 - Event can be related to a second order administrative division (ADM2), such as a district, municipality or commune

4 - Event can be related to a first order administrative division (ADM1), such as a province, state or governorate;

5 - Event can only be specified to a feature that is neither a known point nor a known formal administrative division, but rather a linear feature (e.g. a long river, a border or a road) or a fuzzy polygon without defined borders (informal regions, large radiuses etc.). A representation point is chosen for the feature and employed. Similarly, if a location is only known to be between two points, and these two points are more than 25 km apart, such locations are coded with geoprecision 5.

6 - Event can only be related to the whole country;

7 - Event can only be related to an estimated pair of coordinates at sea or in the air (provided the airplane did not crash as a result of the event; in such cases the location of the crash is coded with the appropriate precision code).

| | | |
|---|---|---|
| **where_prec** | Described above | **integer** |
| where_coordinates | Described above | string(9999) |
| where_description | Described above | text |
| **adm_1** | Described above | string(9999) |
| **adm_2** | Described above | string(9999) |
| **latitude** | Latitude (in decimal degrees) | numeric(9,6) |
| **longitude** | Longitude (in decimal degrees) | numeric(9,6) |
| **geom_wkt** | An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as *well known text* without SRID. | string(9999) |
| **priogrid_gid** | The PRIO-grid cell id (**gid**) in which the event took place. Compatibility with PRIO-grid 1 and 2 (Tollefsen, 2012) is assured.<br><br>**Warning:** We associate every point to the PRIO-grid that contains it, even if the point is in another country than the one officially assigned to the respective PRIO-grid cell through their majority area rule. It is your responsibility to make sure the covariates for the PRIO-grid cell are correct for each event. Further, for the same reason, DO NOT, under any circumstances, first clip out (subset) PRIO-grid by country before merging | integer |

| | | |
|---|---|---|
| | with UCDP GED as data loss will certainly occur. Refer to your copy of the PRIO-grid for further details on PRIO-grid's majority assignment rule (p.3). | |
| **country** | Name of the country in which the event takes place. Note that this variable differs from the country variable in the annual UCDP data, which registers the country of the incompatibility/actor and not the country location of the specific events. | string(999) |
| **region** | Region where the event took place. One of following: {Africa, Americas, Asia, Europe, Middle East} | string(999) |
| **isocc** | The ISO 3-letter code of the country where the event takes place. | string(3) |
| **gwno** | The Gleditsch and Ward numeric identifier of the country where the event takes place | integer |
| **gwab** | The Gleditsch and Ward alphabetic identifier of the country where the event takes place | string(3) |

## 6.5. Clarity

This codes whether the reporting was sufficiently clear for the coder to be able to fully identify the event itself or not.

1 : (denoting high clarity): **events where the reporting allows the coder to identify the event in full**. That is, events where the individual happening is described by the original source in a sufficiently detailed way as to identify individual incidents, i.e. separate activities of fighting in a single location:

Example of such reporting: *"2 people where killed in Banda Aceh town on the 9th of December in fighting between the government and GAM when a car exploded in a main market."*

2 : (denoting lower clarity): for **events where an aggregation of information was already made by the source material that is impossible to undo in the coding process.** The coder merely has access to sources saying that events have taken place (and has aggregated fatality figures), but cannot break apart the reporting into constituent events.

Such events are described by the original source only as aggregates (totals) of multiple separate activities of fighting spanning over a longer period than a single, clearly defined day. Given that the report aggregates multiple incidents into one story impossible to disaggregate back, it is unclear how many battles took place during the time period specified in the source. Thus they are "secondary events", because the form of reporting does not allow the coder to know exactly when the casualties occurred, and how the battles were fought, and the event thus summarises a series of clashes into one event.

Of course, **UCDP has a preference for events with a clarity of 1**; events with a clarity of 2 are just a complement to the former. In fact, often times, it is possible, usually by corroborating multiple reports, to identify some of the clarity-1 events contained in the

description making up the event with clarity of 2. In such cases fatalities in such identified events are subtracted from those given in the clarity-2 event. This leads to clarity-2 events sometimes defying the parameters of the fatality estimates, as the 'high estimate' may at times be lower than the 'best' or 'low' estimate.

Examples of **clarity-2 events**:

"The Ukrainian government informs that 29 people have died in the past six days in a number of clashes with the separatists along the line of conflict".

"in the past 2 months 120 people were killed in operations throughout Assam".

"The responsible for the Aceh military operation indicates that 29 people have been killed in various incidents of fighting over the past five days".

| | | |
|---|---|---|
| event_clarity | **described above** | **integer** |

## 6.6. Time

Each event is defined to have occurred at a certain date. The precision of the dataset is **one calendar day,** starting at 00:00 (midnight) and ending at 23:59 local time.

In many cases, the exact day an event has taken place is impossible to find out with any certainty. In those cases, a temporal precision variable is provided which denotes with what accuracy a specific time period in which the event occurred is known.

The temporal precision variable can have six values:

- 1 – the exact day of the event is know;
- 2 – the exact day of the event is not known, only time period between 2-6 days;
- 3 - the exact day of the event is not known, only the week;
- 4 - the exact day of the event is not known, only the month;
- 5 – the exact day of the event is not known, only the year.

| | | |
|---|---|---|
| **date_prec** | How precise the information is about the date of an event.<br>1: exact date of event is known;<br>2: the date of the event is known only within a 2-6 day range.<br>3: only the week of the event is known<br>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<br>5: the date of the event is known only within a range longer than one month but less than one calendar year. | **integer** |
| **date_start** | The earliest possible date when the event has taken place. | Date YYYY-MM- |

| | | |
|---|---|---|
| | | DD |
| **date_end** | The last possible date when the event has taken place. | Date YYYY-MM-DD |

## 6.7. Fatality figures

This section provides fatality figures for each event.

A note on civilian deaths: Civilian deaths can exist in all three categories of violence.

In state-based and non-state violence, civilian deaths count "collateral" killings, i.e. when one or more civilians are killed as an effect of fighting between the two warring parties. At times, such fighting may even result in only the civilian bystanders receiving fatal injuries. Similarly, imprecise shelling or bombing in the context of an armed conflict is coded as state-based violence unless it is clear (from either reporting or context) that civilians have been explicitly targeted.

In one-sided violence, the targeted and killed civilians are always registered in the `deaths_civilians` column.

| | | |
|---|---|---|
| **deaths_a** | The best estimate of deaths sustained by side a.<br><br>Always 0 for one-sided violence events. | **integer** |
| **deaths_b** | The best estimate of deaths sustained by side b.<br><br>Always 0 for one-sided violence events. | integer |
| **deaths_civilians** | The best estimate of dead civilians in the event.<br><br>For non-state or state-based events, this is the number of collateral damage resulting in fighting between side a and side b. For one-sided violence, it is the number of civilians killed by side a. | integer |
| **deaths_unknown** | The best estimate of deaths of persons of unknown status. | integer |
| **best_est** | The best (most likely) estimate of total fatalities resulting from an event.<br>It is always the sum of **deaths_a, deaths_b, deaths_civilians and deaths_unknown**. | integer |
| **high_est** | The highest reliable estimate of total fatalities | integer |
| **low_est** | The lowest reliable estimate of total fatalities | integer |

### 6.8. Variables present in current versions of GED not present in older versions:

**where_location (GED 1.0 – 1.9)** : renamed to where_description

**coordinate_location (GED 1.0 – 1.9)** : renamed to where_coordinates

**event_type (GED 1.0 – 1.9)**: given the high level of confusion experienced by our users with regards to the actual meaning of the variable, it was replaced by a new concept, **event_clarity**.

**Geocomment (GED 1.0 – 1.9)**: eliminated as a human-legible free-text comment on location names, geocoding sources and alternative spellings proved to have more disadvantages than advantages for data usability in a large, quant-oriented dataset.

**dyad_unique (GED 1.5)**: replaced by dyad_new_id. Dyad_unique was a temporary, stop-gap measure specific to GED 1.5 to prevent an acute problem originating from the merging of three separate systems. The construction of a new UCDP system together with introduction of an UCDP-wide system

**uniq (GED 1.1 – GED 1.5)**: replaced by **id**. Compared to uniq, which was specific to each release of the data, id is persistent, i.e. consistent across releases of the datasets. An entry with the same **id** in version 1.9 describes the same real-life incident in versions 2.0 – 5.0.

# 7. Available Formats (Format Declaration):

The UCDP GED is provided in a variety of formats for use by researchers within different fields and with different needs. All formats are available for download free of charge (no registration required) from the UCDP GED website (http://ucdp.uu.se/ged).

The GED is currently available (as of version 5.0) in the following formats:

**Comma Separated Values (CSV), Excel (XLS), ESRI Shapefile (SHP), R Data Frame (RData) and SQL dumps for use with spatially-aware database solutions.**

Please note that each version comes with attached 'Platform Notices' detailing specifics of the file formats and instructions for loading and using the data.
A brief summary is provided here to help users understand each file format and its compatibility as well as to quickly get started:

**CSV format:** A plain text file containing structured comma separated values. The file is suitable for usage with statistics packages, for processing with various programming languages, etc.

Note that the format implements the **full** CSV specifications as summarized in RFC4180[9]. The full CSV specifications are properly implemented by a large number of

[9] The RFC, summarized, states that data is stored in UTF-8, column names are listed in the first row, lines (rows) are terminated in the Windows new-line system (CR LF), fields (record/cells)

software packages, including *OpenOffice, Stata (using the insheet command or the menu), SPSS, R (using the read.csv function), PHP (using the fgetcsv and fputcsv functions), Python (using the csv module) etc.*

As UCDP uses the defaults, typically absolutely no customization or specifications of additional parameters is needed for the usage of the file[10].

Note that this file does not contain the **geom** variable as it is a plain text file.

**Excel format:** An Excel 2007 compatible file that can be used for visualizing the data in a simple Office-like system.
Note that this file does not contain the **geom** variable as it is a plain text file.

**Rdata format**: An R data-frame. Requires rgdal and suggests sp.

**Esri shapefile format:** An ESRI shapefile package (shp/shx/dbf file) that can be used with various desktop/legacy GIS solutions such as ESRI ArcGIS, QGIS etc. or imported into R (using rgdal) or STATA (using spmap) if the spatial features of the dataset are needed for statistical processing.
Note that this file contains the spatial features (**geom** variable) in the ESRI format.

**Spatial database format:** A dump of an SQL table containing the entire dataset. Importable in compatible SQL databases for processing and analyses as well as usable with new-generation GIS tools (such as QGIS).
Currently supported are Postgresql with the PostGIS extension (versions 8.3 and above with PostGIS 2.0 or above; created on a version 9.1 with PostGIS 2.1) and MySQL (version 5 and above, created with MySQL 5.09) and SQL Server (created with SQL Server 2012). The table is 'flattened', i.e. all the information is stored in a single table rather than in multiple data under a relational model. If you need a more normalized data model, please contact us.

Note that this file contains the spatial features (**geom** variable) in the OGC format (geometry, point, SRID 4326)[11]

**API (machine-to-machine communication):**

An open beta exists for UCDP GED data as a web-service through a RESTful API under the DaaS paradigm. Descriptions on how to use the UCDP GED Beta API are available at **http://ucdp.uu.se/apidocs/**

---

are separated with a comma (','), with each field containing a space, a double apostrophe ("), a comma (,) or a new line (CRLF) being enclosed within double apostrophes ("). Double apostrophes are escaped by double apostrophes ("").

[10] Note that Excel versions prior to MS Office 2016 defaults to Tab-Separated-Values when importing CSV files, and does not support the full specification required by UCDP for the file to correctly load. Please use the provided Excel file. Similarly, ArcGIS has a broken CSV import system as of version 10. Please use the provided shapefile.

[11] The MySQL dump does not contain an OGC binary geometry column, only an OGC Well Known Text column labeled geom_wkt, due to differing implementations of binary geometry in various MySQL engines. The Microsoft SQL Server dump will create the binary blob column on the fly when loading the file.

# 8. Acknowledgements:

The UCDP geo-referencing and event data project is grateful for valuable external input from Halvard Buhaug at the Peace Research Institute, Oslo (PRIO), Nils B. Weidman at Konstanz University, Luc Giradin at ETH Zurich as well as Tomislav Dulic at the Hugo Valentin Centre, Uppsala University, for infrastructural support, ideas and comments.

Current members of the UCDP team, in alphabetic order:

Marie Allansson, Mihai Croicu, Garoun Engström, Erika Forsberg, Helena Grusell, Prof. Håvard Hegre, Stina Högbladh, Gabrielle Lövquist, Prof. Erik Melander (director), Therese Pettersson, Margareta Sollenberg, Ralph Sundberg, Sam Taub, Lotta Themner, Prof. Peter Wallensteen (senior advisor).

# 10. References:

Boschee, E.; J. Lautenschlager, S. O'Brien; S. Shellman; J. Starz, James; M. Ward (2015), "*ICEWS Coded Event Data*", http://dx.doi.org/10.7910/DVN/28075, Harvard Dataverse, V7.

Deborah J. G., P. A. Schrodt, O. Yilmaz (2009). *Conflict and Mediation Event Observations (CAMEO) Codebook*. http://eventdata.psu.edu/data.dir/cameo.html, 2009.

Raleigh, C., A. Linke, H. Hegre, J. Karlsen (2010), Introducing ACLED: An Armed Conflict Location and Event Dataset, *Journal of Peace Research* vol. 47(5):651-660.

Rød, E. G., N. B. Weidmann (2014), Protesting Dictatorship: The Mass Mobilization in Autocracies Database, APSA Annual Meeting in Washington, DC, 2014.

Salehyan, I., C. Hendrix, J. Hamner, C. Case, C. Linebarger, E. Stull, J. Williams (2012). Social conflict in Africa: A new database. *International Interactions*, 38(4):503-511.

Schrodt, P.A., J. Beieler and M. Idris (2013), Three's a Charm? Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. Paper presented at the International Studies Association Annual Convention, Toronto, March 2014.

Schrodt, P.A. Twenty years of the Kansas event data system project. *The Political Methodologist*, 14(1):2–8.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2013). Global Terrorism Database. Retrieved from http://www.start.umd.edu/gtd

Also consult the following UCDP Codebooks:

UCDP Dyadic Dataset Codebook :
http://www.pcr.uu.se/research/ucdp/datasets/ucdp_dyadic_dataset/

UCDP/PRIO Conflict Dataset Codebook :
UCDP Non-State Conflict Dataset Codebook:
UCDP One-Sided Dataset Codebook:
UCDP Actor Dataset Codebook:

**http://ucdp.uu.se/downloads/**

# 9. Appendixes:

## *APPENDIX 1: Temporal Precision Coding and Date Estimation Rules*

This document specifies the qualifications for all temporal precision variable values according to the rules constructed by the UCDP for the GED. It also sets rules for interpretation of time-related expressions and estimation of events' start and end dates. The appendix presents concrete examples that guide temporal precision coding and date estimation procedures.

**Estimation of Start and End Dates**

1. Start and end dates of the events are set according to information in the original sources.

2. Ambiguous time-related expressions (e.g. past few days) are interpreted on the basis of the rules presented below. This ensures uniform estimation of the events' start and end dates throughout the entire GED.

3. If the source does not provide any information about the time period during which the event took place, dates are estimated for three days, counting backwards from the day of reporting and excluding the day of reporting:

    a. "24 rebel soldiers were killed";

    b. "Security forces stepped up operations against the largest insurgent group in Assam state, where a new government was set to take charge on Friday. A police spokesman said four members of the outlawed ULFA were killed in the battles";

    c. "10 bodies found buried in a mass grave in territory controlled by the ULFA rebels".

**Temporal Precision 1 – Daily Precision of Time**

1. If the exact date of an event is known the temporal precision code of 1 is applied. Such events have the same start and end dates that are precisely specified in the news sources either by dates, day names, hours or other specific temporal concepts:

    a. "14th January", "today", "yesterday", "last Tuesday" - date for specified day;

    b. "Monday night" - date for Monday;

    c. "Last night" - date for preceding day of reporting;

    d. "The other day"- date for the preceding day of reporting.

**Temporal Precision 2 – Imprecise Time (2-6 days)**

1. Temporal precision value of 2 should be used in those cases when start and end dates for events are of unspecified character, spanning more than one calendar day though no longer than six days, i.e. shorter than a week:

    a. "Recently", "recent attacks" - dates for 3 days preceding and not including the day of reporting;

    b. "Past/last few days" - dates for 3 days preceding and not including the day of reporting;

    c. "Around 2 July" - dates for three days, 1-3 July, with the stated date +/- one calendar day;

    d. "Over the weekend" - dates for Saturday and Sunday, if source does not include Friday in the concept of weekend and unless specific dates/days for the weekend are provided in the source;

    e. "Since the beginning of the week", "this week" - dates from Monday to the day of reporting;

    f. "Night between Sunday and Monday" - dates for 2 days;

    g. "Past 24 hours" - dates for the day of reporting and the preceding day;

    h. "Past 48 hours" - dates for the day of reporting and 2 preceding days;

    i. "Past 72 hours" - dates for the day of reporting and 2 preceding days;

j. “Past 2 days” - dates for 2 days preceding and not including the day of reporting;

k. “Since Thursday” – dates from Thursday until the day of reporting;

l. “Five-day offensive” - dates for 5 days of fighting including the day of reporting;

m. “Continuous fighting between 13-16 February” - specified dates;

n. “Night-long battle” - dates for 2 days covering the whole night;

o. “Night of clashes” - dates for 2 days covering the whole night;

p. “Last 6 days of January” - dates for 25-30 January, including final date of month;

q. “Late last week” - dates for Friday to Sunday of the preceding week.

**Temporal Precision 3 – Weekly Precision of Time**

1. Temporal precision value of 3 should be used in those cases when start and end dates for events are specified to a certain week, but specific dates are not provided:

    a. “Last week” - dates for Monday-Sunday of the preceding week. Exceptions can be made if there are reasons to believe that the event took place during the week of the reporting (e.g. sometimes “a raid last week” reported on Sunday might refer to the period Monday-Saturday of the same week, then dates for Monday-Saturday of that week should be used);

    b. “Past week” - dates for 7 days including the day of the reporting, unless text indicates that past week refers to an ongoing week (starting Monday);

    c. “First week of August” - dates for August 1-7.

    d. “Week-old offensive” - dates for a week of fighting, 7 days, including the day of reporting;

**Temporal Precision 4 – Monthly Precision of Time**

1. Temporal precision value of 4 should be used in those cases when start and end dates for events are specified to a certain month, but specific dates are not provided:

    a. “Beginning of/early March” – March 1 to March 10/day of reporting;

    b. “Middle of March” – March 15 +/- 5 calendar days, i.e. March 10-20;

    c. “End of/late March” – March 15 to the last day of March/day of reporting;

    d. “A number of weeks”, “recent weeks” - dates for 3 weeks counting backwards from the day of reporting;

    e. “Several weeks” – dates for 3 weeks;

    f. “Earlier this month” – starting the 1st day of the month and ending on the day preceding the day of reporting;

    g. “Last month” - dates for the month preceding the one on which the event was reported;

    h. “A fortnight ago” - dates for preceding 14 days including the day of reporting.

**Temporal Precision 5 – Annual Precision of Time**

1. Temporal precision value of 5 should be used in those cases when start and end dates for events are specified to a certain year, but specific dates are not provided:

    a. “1995” - 1995-01-01 to 1995-12-31;

    b. “Last year” - dates covering the year, YYYY-01-01 to YYYY-12-31;

    c. “Past year” – All dates from the date of reporting back to YYYY-01-01

    d. “Early 1999” – 1999-01-01 to 1999-04-30;

    e. “Mid 1999” – 1999-05-01 to 1999-08-31;

    f. “Late 1999” – 1999-09-01 to 1999-12-31;

    g. “Past 3 months” - dates for 3 months counting backwards from the day of reporting (may not cross over into another calendar year);

h. “Past few months” – dates for 3 months counting backwards from the date of reporting (may not cross over into another calendar year).

## *APPENDIX 2: Geo-precision Coding Rules*

This document gives an overview of the coding rules for geo-precision codes coupled with examples and comments.

**General rules**

1. All geographical locations are coded with moderation with preference given to more certain locations even if they represent a higher level of aggregation over those locations which are less certain but represent a lower level of aggregation.

2. Unclear geographical references with several possible levels of aggregation are coded as the highest possible one. For instance, if there is a town, a district (ADM2) and a province (ADM1) of the same name and the source does not specify to which type of location it refers, then the location will be coded as ADM1.

3. If event location (camp, bridge, road etc.) has the same name as a certain suburb, town or village (e.g. Uppsala IDP camp and Uppsala town), the coordinates for that town or village should be used only if it is known that the event location is within or close to (within 25 km) that town or village. If information about the locations' proximity to that town or village is not available, the location is aggregated to the lowest available administrative division. For instance, if it is not known that Uppsala IDP camp is within 25 km from Uppsala town, coordinates for Uppsala municipality (ADM2) should be used.

4. If the source refers to a certain location (e.g. river, forest, lake, park, mountains etc.) that is not similar in size with a locality, or that is not a point, a representation point is created with precision 5. If that location lies within an ADM2 or ADM1, the ADM2 or ADM1 is attached to the representation point. Do not aggregate e.g. rivers or national parks to administrative divisions if representation points can be made.

5. When coding historical observations the GED uses the names of the administrative divisions in force at the time of the reporting. If the boundaries of ADM1 have changed over time in a country, the dataset uses estimated coordinates for older provinces based on the relevant seat of the ADM1 at the time of the event.

   A history of administrative changes is tracked internally by the UCDP system in a data structure referred to as a geotree. If you require access to such files, contact us.

**Geo-precision 1**

Geo-precision value of 1 is used if the location information corresponds exactly to the geographical coordinates available. Each pair of coordinates is also coupled with names for ADM1 and ADM2 when available.

1. “City”, “town”, “village” , “location”, “locality”- centroid point coordinates;
2. “District”, ”quarter”, ”neighbourhood”, “locality” (of town) - coordinates for town centroid point are applied here, and not the specific section of it, though the name and details are kept in text in parenthesis in “Where";
3. Air battles if location is clear, i.e. “a plane was shot down over Kitgum”.

**Geo-precision 2**

If the location information refers to a limited area around a specified location, coordinates for that location together with the geo-precision value of 2 are used.

1. “Near/in the vicinity of/adjacent to/just outside/around Kitgum town” – coordinates for Kitgum town;
2. “Pietermaritzburg area” – coordinates for Pietermaritzburg town;
3. “Outskirts/suburbs of Bujumbura city” – since outskirts and suburbs are understood as relatively independent and distant entities coordinates for Bujumbura city should be used;
4. “17 km from Uppsala town” – if the event takes place within a distance of 25 km from a specified location, coordinates for that specified location are used;
5. “North of Luanda city”, “southeast of Y mountain” - unspecified distances from a specified location are understood to be near the stated location;
6. “Bujumbura city towards Gishingano village” – if coordinates for Gishingano village can not be retrieved, then coordinates for Bujumbura city will be used;
7. “Niuland village near Dimapur town” - if coordinates for Niuland village are not available, but coordinates for Dimapur town exist, the latter are used;
8. “Dungu territory in DRC” – third level administrative divisions (ADM3), if small enough to have an approximate radius of 25 km or less, receive a precision code of 2.

**Geo-precision 3**

If the source refers to or can be specified to a larger location at the level of second order administrative divisions (ADM2), such as district or municipality,

the GED uses centroid point coordinates for that ADM2. If these are not available, representation coordinates for a town within that ADM2 are used. The name of the ADM2 in force at the time of reporting is recorded in the variable ADM2.

1. "Arusha district, Arusha province" - coordinates for Arusha district (ADM2);
2. "Burambi commune, Burundi" – coordinates for Burambi commune (ADM2);
3. Air battles if unclear location - if the battle takes place "over" a certain ADM2, coordinates for that ADM2 will be used;

**Geo-precision 4**

If the location information refers to a first order administrative division, such as a province (ADM1), the GED uses the coordinates for the centroid point of ADM1.

1. "Cibitoke province, Burundi" – coordinates for Cibitoke province (ADM1);
2. Air battles if unclear location - if the battle takes place "over" a certain ADM1, coordinates for that ADM1 are used;
3. If the ADM2 in which the event took place in unclear (e.g. different sources refer to different ADM2s in which the same event took place), the location is aggregated to the ADM1 level;

**Geo-precision 5**

Geo-precision value of 5 is used in these cases:

1. If the location information refers to parts of a country which are larger than ADM1, but smaller than the entire country such as "Southern Lebanon", "Northern Uganda". In these cases, a representation point is created for that part of the country and used as a representation of that area together with geo-precision value of 5. Note that these points are stored and reused consistently by the UCDP (thus, all events assigned to "Northern DR Congo" will have the same coordinates recorded).

2. If a pair of coordinates is estimated as a representation point for a linear, non-administrative polygon or fuzzy geographic feature (river, informal area, large lake etc.). For example, if the location is on the border between two countries and the location of such point is not precisely known, a pair of estimated coordinates will be used together with geo-precision value of 5. For example, "on the border between Uganda and Sudan" will be coded as "Uganda/Sudan border" with the coordinates for a selected point on the border between Uganda and Sudan; Note that these points are stored and reused consistently by the UCDP (thus, all events assigned to "Uganda/Sudan border" will have the same coordinates recorded).

3. If the location information refers to islands which are not an ADM1 or 2 of their own. For example, “Zanzibar island” will be understood as eastern part of Tanzania and receive geo-precision value of 5. If a pair of coordinates for that island is not available in the gazetteers, it can be represented by an ADM1 in that island.

4. If the location is not specific and need to be estimated (for example, “road between Pader and Kitgum”, “along Aswa river” etc.), or the location is more than 25 km away from another location (for example, 75 km south of Kitgum town), then a representation point is created for that point. This is done even if the two points are located in the same ADM2 As such, if an event is described as taking place on “the road between Yei and Rasul in Yei district of Equatoria State”, then a point is estimated on that road, with precision 5, with both the ADM1 (Equatoria state) and the ADM2 (Yei district) coded.

**Geo-precision 6**

If the location information refers to an entire country, centroid point coordinates of that country are used. Also, if the location is not provided/is unclear/refers to several locations which can not be split and covers the whole country and a particular activity area of the actor is not clear, centroid point coordinates of that country are used.

1. "Germany" - centroid point coordinates;

**Geo-precision 7**

If the event takes place over water or in international airspace, the geographical coordinates in the dataset either represent the centroid point of a certain water area or estimated coordinates according to similar techniques as presented above for geo-precision code 5.

For air events, precision code 7 is used only if the death is not the effect of or did not result in the airplane crashing (in such a case, 1-5 precision codes are used with the location of the crash).

1. “Southern ocean” – centroid point coordinates;

2. “Bay of Bengal” – centroid point coordinates;

3. “37 km off the coast from Stockholm city” – estimated coordinates for a point 37 km and 90 degrees off the coast of Stockholm.

4. “the minister was stabbed on an airplane en route to Delhi after departing Islamabad” – coordinates for Islamabad airport, precision code 7.