Political Instability Task Force Atrocities Event Data Collection Codebook *

Philip A. Schrodt and Jay Ulfelder
Point of contact: schrodt735@gmail.com

Version 1.1b1 : September 12, 2016

*This work is sponsored by the Political Instability Task Force (PITF), which is funded by the Central Intelligence Agency. All of this work and exposition is the responsibility of the authors alone and does not represent the views of the US Government.

1 Overview

This document describes the motivations, definitions, and coding procedures used to create the Political Instability Task Force (PITF) global dataset on atrocities, by which we broadly mean the deliberate killing of non-combatant civilians in the context of a wider political conflict. This codebook—originally prepared by Jay Ulfelder in 2006 as a revision of an earlier document produced in collaboration by Jay Ulfelder and Philip Schrodt in consultation with the project's sponsors, then extensively revised by Philip Schrodt in 2016—attempts to incorporate comments made by several experts who reviewed the design and results of an experimental data-collection effort. That experimental effort was undertaken in autumn 2003 and led by Philip Schrodt, University of Kansas, under subcontract to Science Applications International Corp. (SAIC). The original SAIC project management was provided by Irwin Jacobs, Program Manager, and Jay Ulfelder, Technical Manager. Additional text for the codebook has been provided by Dennis Hermrick, Milos Jekic and Taylor Price, University of Kansas.

The following narrative describing the data is included at the request of the sponsor:

> "The Political Instability Task Force (PITF) Worldwide Atrocities Dataset is a global dataset that describes, in quantitative terms, the deliberate killing of non-combatant civilians in the context of a wider political conflict. This data collection project, which is still ongoing, is intended to advance efforts to understand and anticipate atrocities, i.e., the deliberate use of lethal violence against non-combatant civilians by actors engaged in a wider political or military conflict. The practical objective of this project is to create a dataset representing a reasonably systematic sample of atrocities occurring worldwide in recent decades in order to: (1) enable the development of statistical models that might be used to identify countries vulnerable to the occurrence of atrocities or, if atrocities are already occurring, to an escalation in their rate or intensity; and (2) create a descriptive record that might be used by researchers with an interest in particular countries or conflicts. The effective date of data in this dataset is 1 January 1995 to the present date. Data are updated monthly.
>
> "This data-collection is sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency. The data set is the responsibility of the authors' alone and does not represent the views of the US Government."

The data described in this codebook are available at:

`http://eventdata.parusanalytics.com/data.dir/atrocities.html`

Comments on the Version 1.1b Update

This version of the codebook is a very substantial and long-overdue revision reflecting changes that have occurred in the data collection protocols since the project began over a decade ago. Particular attention has been paid to explicitly indicating coding decisions that may account for differences between these data and similar sets, for example the UCDP one-sided violence dataset[1] or the datasets of the Human Rights Data Analysis Group.[2] The document also includes an appendix explaining why a few of the variables found in the original codebook have been dropped.

[1] `http://www.pcr.uu.se/research/ucdp/datasets/ucdp_one-sided_violence_dataset/`

[2] `https://hrdag.org/`

This document is both a conventional codebook but as an extended guide to some of the ambiguities in the data: this could be useful both in explaining some outliers, and also may provide useful advice for projects attempting to code similar data. The word "I" in the revisions refers to Schrodt. In keeping with its "work-in-progress" status, the tense used to describe the coding process is variously past, present, and future: sorry.

2 Project Motivations

This project is intended to advance efforts to understand and anticipate atrocities, which we understand as the deliberate use of lethal violence against non-combatant civilians by actors engaged in a wider political or military conflict. Conflicts producing atrocities may be motivated by a variety of interests: not only the desire to wield state authority, but also the desire to control economic resources, to change the status of a particular ethnic or religious community, or to defeat another state in an international war.[3]

The practical objective of this project is to create a dataset representing a reasonably systematic sample of atrocities occurring worldwide in both in recent decades and updated on a monthly basis in order to: 1) enable the development of statistical forecasting models that might be used to identify countries vulnerable to the occurrence of atrocities or, if atrocities are already occurring, to an escalation in their rate or intensity; and 2) create a descriptive record that might be used by researchers with an interest in particular countries or conflicts.

To our knowledge, most existing datasets on atrocities either cast the "macro" event of a larger conflict process as the unit of observation—for example, Harff's (2003) or Fein's (1993) lists of genocides or Rummel's list of democides[4]—or focus on "micro" events within the context of a single conflict process—for example, Ball, Kobrak, and Spirer's (1999) dataset listing instances of human rights violations in Guatemala, 1960-1996. While many of these datasets are important and useful, they limit researchers to analyzing either the onset of conflict processes involving atrocities—and often more specifically to conflict episodes that rise to the level of genocide—or to analyzing the dynamics of violence in the context of a genocide. They do not allow researchers to draw inferences about the likelihood of any deliberate violence against non-combatants in situations where none has yet occurred, or the dynamics of such violence in situations that may or may not rise to the level of genocide—both of which may of substantial concern to policy-makers, researchers, and other interested parties.

To enable the development of models of the dynamics of violence applicable to countries that may or may not experience atrocities on a massive scale, we aim to record information from press accounts of atrocities occurring in all countries of the world during our period of observation. The data set that results from this global approach will allow us to study the risk of atrocities that may or may not accumulate to the level of genocide or "democide" while also avoiding the problem of selection on the dependent variable that would contaminate any effort to pool observations from existing atrocity event data sets.

[3] We do not intend to suggest that each atrocity is motivated solely or even partly by the objectives at stake in a wider conflict. We are simply noting that atrocities occur in the context of virtually every form of violent political struggle.

[4] `https://www.hawaii.edu/powerkills/NOTE5.HTM`

We are primarily interested in situations where state or non-state actors use violence against non-combatant civilians as a tactic in wider conflicts. We also recognize, however, that the unambiguous intent to kill non-combatants is often difficult to establish, and that such gross violations of human rights sometimes occur in the context of conflicts between or among non-state actors, or in situations where state authority has decayed so completely that the state vs. non-state archetype does not apply. Because violence is often very localized, and in some large states such as India, Nigeria and Indonesia, multiple conflicts may exist that are essentially independent of each other, we intend to geolocate every incident, and to the extent possible from the reports, identify the political, ethnic or religious groups that the victims are associated with.

In our view, the ambiguity of many situations involving attacks on non-combatant civilians necessitates a flexible data collection scheme. In the approach described below, we have attempted to avoid asking coders to base the decision to include or exclude reports on the basis of their own assumptions or inferences about perpetrator motivations and identity. Instead, we have attempted to cast our net as widely as possible, recording information on all events that, according to the description offered by journalists, might be interpreted as representing the deliberate use of lethal violence against civilians as part of a political conflict.

By capturing as many potentially relevant events as possible and explicitly recording their ambiguities, we intend to allow researchers to employ various approaches when analyzing these data, ranging from narrow interpretations of what constitutes an atrocity to broader ones, or from restrictive rules about the character of reporting on these events to looser ones. To our knowledge, such an effort systematically to document atrocities worldwide has not been undertaken previously.

By focusing our research on reports of killings or series of killings, we aim to treat the use of violence against civilians in any political conflict as a variable, open to empirical study. This approach recognizes that such killings may occur before or after a large-scale conflict is conventionally understood to have begun or ended; such killings may also occur in or outside the context of a conflict process considered genocide by other researchers. We believe this strategy will allow us to study not just the occurrence but also the dynamics of mass atrocities in the widest variety of contexts.

3 Defining the sample of stories

This data set can be thought of as a conservative but precise set of cases. All of the following criteria need to be met for a case to be included:

1. In the Factiva database, incident is found in the "Headline or lead paragraph" using the following search

```
(killed or kills or massacre or bomb or bombing or bomber or beheaded or mass
grave) not (crash or crashed or accident or accidents or funeral or flood$
or house fire or apartment fire or lightning or mine blast or lion or crocodile
or rhino$ or elephant$ or hippo or shark or disease$ or Ebola or cholera or
drone strike or drone attack or suspected drone or murder-suicide or hit-and-run
or half-staff or tornado or tornadoes or cyclone or landslide or landslides
or mudslide or avalanche or earthquake or quake or typhoon or volcano or dengue
or zika)
```

and the English-language reports of the following sources:

- Reuters News
- CNN
- New York Times
- Agence France Press
- BBC
- Associated Press
- All-Africa (after 2013)

2. At least five non-combatants have been killed in an incident or the incident is a "targeted killing"
3. The date of the incident can be determined to within a week, more or less
4. The location can be determined at least to roughly the level of a province
5. The incident does not involve the United States as either target or perpetrator

3.1 Discussion of the inclusion criteria

These inclusion criteria—in addition to some definitional differences—should explain most of the differences between the this data set and other data on lethal violence against civilians such as those of the UCDP and the reports of various human rights organizations. Most of these were motivated by one or more of the following issues

- The focus, discussed extensively in Section 2, on monitoring possible precursors to develop indicators for monitoring and developing statistical models for anticipating large-scale atrocities against civilians
- The requirement that the data—which is open-ended—be available in near-real-time: it is updated monthly.[5]
- The relatively limited resources available to the project and the fact that the coding is done entirely manually, albeit with the assistance of some machine-based tools for organizing the information and a highly optimized web-based coding form.

3.1.1 Search string and source list

The search string was developed in the initial phases of the project and has been maintained consistently except for the addition of a few terms (for example "beheading", "Ebola" and "zika") that only began to show up in more recent phases of the project. We originally experimented with a number of more specific phrases instead of the core "killed or kills" but the general approach works best. This currently picks up between 3,000 and 4,000 articles a month, with typically about 5% to 10% of these being relevant, those numbers of course fluctuating depending on the level of violence in various parts of the world. Except for the initial search string, the determination of whether a story is relevant is determined on a case-by-case basis by coders. This is usually

[5]Prior to 2016, the publicly-available data were embargoed for three months: this restriction was removed in January 2016 and the data are now posted approximately two weeks after the month for which they have been collected.

unambiguous, with most of the rejected cases involving combatant deaths, noncombatant deaths below the threshold, and the word "killed" used in a metaphorical context ("The show was killed after only a single season..."). When multiple reports are available for an incident, all will be included in the `Citations` field except when one gets a situation where a high-casualty incident occurs in an otherwise peaceful major city (e.g. Paris, Brussels, Istanbul, Abuja, Bangkok), which can generate literally hundreds of reports, in which case only a sample will be cited and this fact will be noted in the `Comments` field.

The source list includes the two major international sources Reuters and Agence France Press,[6] as well as less-frequent Associated Press. BBC is primarily useful for the monitoring it does of broadcasts and web sites, providing an extensive secondary source of local reports. The same is true of All-Africa, which has been tremendously useful for monitoring Nigeria and the eastern Democratic Republic of Congo. CNN rarely provides much beyond the other sources, and the now-cash-strapped *The New York Times* is basically useless. Were this project to start over, Xinhua probably should be added to the list, and possibly the English service of al-Jazerra.

Keep in mind that Reuters and AFP make *extensive* use of local "stringers" who are usually journalists from major city newspapers, so while these sources are "international" (and in English), the reporters are usually locals fluent in one or more local languages, not just some character out of a Graham Greene novel hanging out in a ratty hotel bar near the national palace in a whiskey-induced haze waiting instructions from their distant and detached editor to cover The Big Story.[7] Usually.

At some point fairly early in the project, we did a systematic assessment on using whatever the current iteration was of the Open Source Enterprise,[8] which changes its name at about the same frequency most people replace their personal vehicles, and a decision was made that the marginal contribution of cases did not justify including it. That situation may or may not still be true.

3.1.2 Five death threshold for non-combatants

We have only coded incidents involving five or more non-combatant deaths. We attempted at one point to lower this threshold to "one" but the data collection demands proved completely overwhelming, as this involved assessing every murder and ambiguous accidental death reported anywhere in the world in the international media. "Five" has no underlying theoretical justification; it merely provides a threshold above which we can confidently code all of the reported events given our available resources, since in most parts of the world this is actually a fairly high threshold.[9]

The criteria by which an individual is judged a "non-combatant" or a "targeted killing" are discussed in detail in Section 4 but note in particular that they exclude both military and civil security forces.

[6] Despite the name, AFP has been private for about two decades.

[7] Nor, in my experience in the Middle East, are they predominantly male.

[8] `https://www.opensource.gov/`

[9] This is not true for the U.S. (which, again, we aren't coding)—a *Washington Post* study found killings of 4 or more people on average occurred about once a day in the U.S. in 2015—and such events will be fairly common in active war zones, but otherwise this threshold eliminates most instances of private violence. Reference:
`https://www.washingtonpost.com/news/wonk/wp/2015/11/30/there-have-been-334-days-and-351-mass-shootings-so-far-this-year/`

3.1.3 Date

For most incidents, the date is unambiguous to within a day. As discussed in Section 5, the data set excludes the reports of large number of deaths over a long period of time which are often found in IGO and NGO reports. This limitation also means that victims found in mass graves are not recorded unless the deaths can be linked to a specific date or limited period, which will occasionally occur (as for example in the mass graves associated with the ISIS massacre of of Yazidi men in the village of Khocho, south of Sinjar, on 15 August 2014). The limitation of timing within a week is not sharply enforced—as always, exceptions will be noted in the `Comments` field—but does exclude vague periods such as "last month."

3.1.4 Location

As with the restrictions on dates, these data are restricted to incidents that can be associated with a specific location. In most cases, this can be resolved to a town or village, though ambiguity to the level of a provence is accepted. We usually do not try to resolve locations beyond the city level, even when detailed information is given (e.g. "Tahrir Square"); city coordinates are those provided in `www.geonames.com`.

The data do not include cases where the totals are provided only at a national level. Similarly, situations where deaths have occurred but the location is unclear—typically bodies found floating down a river—are not coded.

3.1.5 Exclusion of the US

The exclusion of incidents involving the United States as either target or perpetrator is due to legal limitations on the Central Intelligence Agency, which is funding this collection. These limitations date to the 1970s—see `https://en.wikipedia.org/wiki/Church_Committee`—and were intended to limit the ability of the CIA to monitor (or analyze) the behavior of U.S. citizens. While open-source data collections such as this were probably not the intent of the original legislation, the PITF has interpreted those restrictions cautiously and therefore such episodes are not included in this data set.

4 Defining and Coding Atrocity Events

For purposes of this project, we define an atrocity as implicitly or explicitly political, direct, and deliberate violent action resulting in the death of noncombatant civilians.

The victims' status as noncombatant civilians is a critical element of virtually all conceptions of an atrocity. This criterion focuses our research on the use of violence against individuals not intentionally involved in an armed struggle. Following Valentino, Huth, and Balch-Lindsay (2001: 8), we define a noncombatant civilian as "any unarmed individual who is not a member of a professional or guerrilla military group and who does not actively participate in hostilities by intending to cause physical harm to enemy personnel or property." Further following those authors, we note that associating with combatants, feeding or sheltering them or participating in non-violent political activities in support of combatants does not transform a civilian into a combatant.

Our conception of noncombatant civilians, however, does not extend to individuals sentenced to death with due process by a state's criminal justice system, nor does it include individuals killed while engaged in acts of violent crime, such as rioting, brawling, or looting where only the individuals engaged in those activities were targeted (but see the discussion of the "Clash" category in Section 11.1). It is also worth noting that our definition does not depend on the racial, ethnic, religious, or political status of the victims—only their status as noncombatants. In this manner, we cast our net more broadly than researchers interested in the narrower issue of genocide.

Of course, claims about the combat status of killing victims are often contested, and journalistic accounts of potential or apparent atrocities sometimes include the claims and counterclaims of interested parties. This project adopts a skeptical stance on claims about the combat status of the victims. Consistent with our broader attempt to cast the net as widely as possible, if some parties claim the victims were combatants or other legitimate military targets but those claims are contested, we include that event in our dataset and record the fact that there were competing claims about the victims' status.

We also err on the side of inclusion by assuming that when news sources do not apply a military or paramilitary label to victims of killings, those victims were most likely non-combatant civilians and the event should be included. Again, however, we attempt to capture this ambiguity in our database with a field indicating whether the victims were explicitly described as non-combatant civilians.

Finally, departing somewhat from the Valentino, Huth, and Balch-Lindsay definition, we include as "non-combatants" individuals who may be combatants in a different context, but who at the time they were killed were unarmed and unable to defend themselves. This would include, for example, members of guerrilla groups who had come into refugee camps unarmed in order to get food or medical care, or off-duty police eating at a pizza stand. This does not include situations where the individuals were in a military or para-military situation (for example a check-point, recruiting station, guerrilla base or police barracks) which the attackers could reasonably have expected to be defended, even if the individuals happened to be unarmed (e.g. sleeping or still unarmed in the case of recruits) at the time of the time of the attack.[10] If there is doubt about whether the victims qualify as "temporary noncombatants," we code the case but indicate the ambiguity in the `Comments` field.

The modifiers direct and violent focus our research on situations where deeds committed by the perpetrators are the proximate cause of noncombatant civilian deaths. These characteristics distinguish atrocities from other actions considered war crimes or crimes against humanity in which individuals or organizations make decisions that put non-combatants at grave risk but do not directly kill them. Such situations include attacks on humanitarian aid missions; the use of civilian facilities, such as hospitals or schools, as "hiding places" in the context of an armed conflict; and efforts to restrict or destroy civilian supplies of food or water.

The references to action that is deliberate and implicitly or explicitly political focuses our research on violence intentionally perpetrated to political ends, as opposed to accidents, acts of nature, or acts of private or personal violence. This criterion gets to the intentions and agency of the perpetrators, aspects that are notoriously difficult to establish with certainty even in a court of

[10]The killing of individuals at recruiting stations is a particularly difficult case because these have become fairly common targets for suicide bombings. tend to have high casualties, and most of the victims are unarmed. On the other hand, a recruiting station has a clear military purpose and the individuals are presumably intending to become combatants, and thus is quite different than a similar attack on a marketplace.

law, let alone from brief journalistic accounts. To address this dilemma, we again err on the side of inclusion by excluding only those events that appear to fit decisively in one of several categories in which we are not interested. We are therefore excluding

Acts of Nature: Any event in which deaths are not the result of human actions, e.g., natural disasters, falling trees, wild animal attacks, etc.

Accidents: Any happening that is not intended or expected. Apparently accidental civilian deaths resulting from military or police action are included but coded as **Noncombatants not intentionally targeted** in the `Perpetrator intent` field.

Suicide: Multiple victims kill themselves and no one else is harmed. Mass suicides are very rare, though if there is evidence that some individuals, for example children, were killed without their consent these would be coded.[11]

Private Murders: A person or persons are killed over a personal matter or in the course of a criminal act with no apparent political intent. The key issue is whether the perpetrator or perpetrators apparently acted on behalf of a state agency or a communal or political group.

When there is ambiguity about whether an event should be included, it will be coded but the reservations will be recorded in the `Comments` field.

In general, something is political when it is intended (realistically or otherwise) to have an effect on some collective group that goes beyond the group engaged in the action. That collective group frequently involves a government, but it can also be an ethnic or religious group or any other distinct group of individuals. Criminal activity, in contrast, only benefits the individuals who are involved in it and is not intended to have a collective impact.

The two obvious "grey areas" in this definition are:

1. When criminal groups engage in "Robin Hood" activities where they attack one group for the supposed benefit of another group (stereotypically the rich for the benefit of the poor, though in ethnic conflicts this may become very fuzzy); well-organized criminal groups may also provide re-distribution through the provision of some social services (distribution of food; support for religious institutions) in populations that support them.

2. Collective action by representatives of one social class (typically landowners or factory owners, or rather the thugs and death squads hired by them) against another social class (typically workers or peasants), where there is a direct benefit only to the wealthier class.

The ambiguity occurs because in many situations, one gets a mix of collective and individual benefits, so there is no single distinguishing characteristic. So for example poor farmers in one ethnic group might attack even poorer farmers in a different ethnic group—as happened in South Africa near the end of the apartheid era, and also is now a frequent occurrence in clashes between farmers and herders—and such actions have both a collective and individual effect. Any time there is some indication that the motivation for the violence might be interpreted as being political, assume that it is, but indicate any reservations about this in the comments.

[11] This also is problematic in some suicide bombings where children or individuals of limited mental capacity are carrying the explosives, as well as the not-infrequent situation where a suicide bomber changes their mind (based on reports from eyewitnesses) but the explosives are set off by remote control. Because it is difficult to consistently resolve these cases, the general rule that perpetrators are not included the casualty counts is still applied.

4.1 Special Case: Targeted Killings

Beginning in 2013, we have consistently been coding "targeted killings" of elites and community leaders even when only a single person is killed. The logic of this is that these either could be a low-level precursor to expanded violence, an attempt to intimidate a community into either flight or silence, an attempt to remove the leadership of the community, or an effort to reduce the quality of life of a community. Or all of these.

The individuals who are consistently coded in the category are

- Any formal or traditional community leaders
- Government officials not associated with security: this excludes for example judges, prosecutors, and police officials.
- Journalists
- Activists in NGOs dealing, for example, with environmental, ethnic, or women's rights
- Religious leaders
- Union leaders and organizers
- Medical personnel
- Educators

While most targeted killings are quite unambiguous, with the victim assassinated in mid-day with no attempt at robbery, in some instances it is unclear whether the intent was political rather than criminal, either a street crime or, as we found in a report of a killing of a local official in the Philippines in August-2015 "the killing could be connected to management infighting at the farming cooperative." Entertainers (e.g. Mexican musicians, who seem to disproportionately incur the wrath of drug lords) are coded only if there seemed to be a clear political or ethnic focus to their music; this of course is a really subjective call. As with all ambiguous reports, these will be coded and reservations will be recorded in the `Comments` field.

5 Event Type and Reporting

5.0.1 Event Type

In general, our intention is to code incidents resolved to the level of a locality, so if a single article provides information on multiple events, a separate record is created for each event.

Incident An atrocity perpetrated by members of a single organization or communal group, or by members of multiple organizations or groups reportedly acting in concert, in a single locality within a 24-hour period.

By locality, we mean a single village, town, or city. We recognize that villages, towns, and cities are not entirely comparable levels of analysis; for example, the distance between two neighborhoods in a single city may be greater than the distance between two rural villages. Even so, we believe the act of traveling from one village or town to another represents a kind of discontinuity that often requires perpetrators to commit anew to their behavior and thus represents a break point

between discrete events. See also the discussion in Section 7.6 on the ambiguities of local names: it is frequently unclear whether attacks on multiple "villages" involve actions separated by some distance, or merely, say, crossing a road.

Campaign. A set of reportedly related atrocities perpetrated by members of a single organization or group, or by members of multiple organizations or groups reportedly acting in concert, over multiple days in the same locality, on the same day in multiple locations or in the case of cities, since we only resolve locations to the city level, the same day in the same urban area.

Even under the looser definition of a campaign, the event—not the report—is the unit of observation. Where a single report provides sufficient information to distinguish multiple incidents in either time or location, each of those incidents is recorded separately.

5.1 Comment on the "incident" vs. "campaign" distinction

With the hindsight of a number of years of coding experience, this turns out not to be particularly useful distinction, and as much as anything it depends on the substance of the report and—particularly—the size of the locality. More generally, because we do not have access to the plans and motivations of the perpetrators, we have no consistent way of knowing whether incidents are connected or not, and it is very clear that in some cases incidents that are widely separated geographically were coordinated. In informal discussions we've had with other groups collecting similar data, everyone seems to encounter this same problem: "campaign" looks like a useful category in theory, but in practice it becomes completely ambiguous.

In the early phases of the project, we coded reports—usually from NGOs and IGOs—of "campaigns" that aggregated across long time periods and large geographical areas, typically of the form "over the past six months, <an appalling large number> of people have been killed in violence in <a place you probably don't want to visit right now>." These have been dropped for at least three reasons:

- These reports are usually not available on a timely basis, and often are only available several months after the events occurred
- The reports are necessarily secondary, and the methodology by which the numbers are derived is considerably less clear than the incident-level reports which constitute the bulk of the data
- From the standpoint of statistical modeling, these numbers become outliers—both in terms of their magnitude and the time covered—and generally need to be removed from the analysis anyway

NGOs and IGO reports of individual incidents that can be localized in time and space are coded even if the reports are only available some time after the event: this is often the case when the investigation involves considerable fieldwork and interviewing of witnesses.

5.2 Campaign Identifier

This is no longer used.

5.3 Event Report

This second variable here is used to indicate the nature of the reporting on the event.

Eyewitness Account, Not Contested: The report contains details, usually attributed to a specific source, consistent with the reporting having talked with someone who witnessed the event in question.[12]

Secondary Account, Not Contested: Press reports offer a second-hand account of an event that is represented as a fact. Press accounts of reports by government agencies, international organizations, and non-governmental organizations (including human rights groups) describing atrocities should be coded in this category.

Rumor/Allegation: Journalistic accounts report an event as a rumor or an unverified allegation. Use this only if all sources indicate that the incident was a rumor. This category is used only rarely.

Contested: This category is no longer used: see the discussion in 13.6.

6 Event Date

Record the reported start and end dates of each event as accurately as possible. In cases where the event comes to light later—for example, because a mass grave is discovered—record the reported date of the occurrence of the atrocity as the time of the event, not the date of the related discovery (though it may be useful for other purposes to keep a record of the timing of the discovery as well; this should go in the `Comments` field). For incidents, record the event date in the `Start Date` fields and use 9s (i.e., 99, 99, 9999) in the `End Date` fields. For campaigns, record information in both sets of fields.

Dates are typically unambiguous except when an attack occurs overnight and the time of the start of the attack is unclear: in those instances use the second day (that is, the day after midnight). If an incident can be resolved to within a week but is otherwise unclear, use a date in the middle of the week and note the ambiguity in the `Comments` field.

7 Event Location

Like a number of contemporary conflict datasets, every event has been geolocated, usually to a level equivalent to a few kilometers. Both geographical coordinates and location names are provided.

Geolocation is the most labor-intensive part of the coding process. Once the name of a location has been established, the following sources (in order) are used to try to geolocate it:

- `http://www.geonames.org`: This is the primary source for coordinates. If something can't be found at the "Locality" level, it is noted in the `Comments` field

[12] This is a broader criteria than was used in the early stages of the project, where we wanted the journalist to have witnessed the event or its aftermath, and recognizes the now wide prevalence of "cell phone journalism" where witnesses are interviewed at a distance.

- maps.google.com: Use the Ctrl-click/"What is here?" to get general coordinates for a place on the map; these can be automatically converted to degrees-minutes-seconds by the coding software.
- Wikipedia: This has a few places that geonames.org does not have—this is currently true of Somalia in particular—and sometimes geonames.org will link to this. The French Wikipedia is quite useful for locations in Francophone Africa: Google searches will locate these entries.
- Use Google to find other reports of the incident which have more (or differently transliterated) information on the location; if additional sources are used, provide their URL in the Comments field.

7.1 Country

This contains the ISO-3166-alpha-3 code for the country:
http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3

7.2 Region

This is the province or "1st order administrative division" from www.geonames.org

7.3 District

This is generally the "2nd order administrative division" from www.geonames.org when this information is available. It is usually blank for urban areas.

7.4 Locality

For urban areas, this is the name of the city. In rural areas, this is the name of the smallest geographical area identified in the story. If that location cannot be geolocated, use the District and indicate this in the Comments field (except for Nigeria, where geocoordinates for rural locations always refer to the "Local government area" in the District field.)

7.5 Establishing coordinates

If the name matches something in www.geonames.org, just copy that coordinate. If it does not, or it is ambiguous, you are headed down a rabbit hole. A few common contingencies

- It is sometimes possible to guess alternative transliterations: if one of these works, use enter the version that generated the successful www.geonames.org hit.
- "Every culture has a 'San Jose' "—that is, a place-name that is so common there may be dozens of instances even in a single province. Enter one of them and record the ambiguity in the Comments field.

- Use Google maps to select an arbitrary spot corresponding to descriptions such as "20 km north of..."[13] and "near the Rwandan border" and record the ambiguity in the `Comments` field.
- Use Google to find other reports of the incident which have additional (or differently transliterated) information on the location; if additional sources are used, provide their URL in the `Comments` field.
- `maps.google.com` has a very quirky habit of displaying different place names (and languages) depending on the level of magnification, and experimenting with this can be helpful if you know the general vicinity of a place name but don't see it initially. In a few instances, the "What's here" option will display a place name not visible anywhere on the map.

7.6 Additional comments on geolocations

1. Events outside of cities in Nigeria can almost always be resolved in the level of the "local government area" but not to the level at a "village."

2. `www.geonames.org` and `maps.google.com` coverage is very poor in the Central African Republic, eastern Congo, western Sudan and almost nonexistent in South Sudan. I've been supplementing the Darfur region of Sudan with some maps in PDF files from UN agencies but even these are sometimes insufficient.

3. Transliteration is a particular issue in southern Somalia where two or three systems seem to be in use, and in a number of cases the place names are being transcribed from radio reports or phone calls.

4. The phrase "village" is quite ambiguous and in West Africa seems to refer more to a neighborhood than to the distinct small separated habitations one visualizes from "villages" in England, France or Germany. These place-names also rarely recorded in the `www.geonames.org` database, though sometimes they will be found on `maps.google.com` at some levels of magnification. Situations become even more ambiguous on the edges of urban areas: for example the `maps.google.com` "Earthview" photos along the N2 highway north of Goma, DRC show extended areas of almost continuous low-level settlement following the road: there are apparently local names for some of these areas but we rarely have access to this, and in cases where the details of a report come via cell-phone conversations with survivors, the reporter is simply conveying whatever names their source was using.

8 Perpetrators

These variables record the reported identity of the event's perpetrators and their reported relationship to the state. Code all of the groups that were explicitly reported as responsible; for example, if both the police and paramilitary forces were reported as responsible, enter codes in both of these fields. In all instances, any group names and keywords indicating something about the perpetrator's identity (e.g., "ethnic militia" or "partisan thugs") should be recorded in the `Description` field.

[13] This procedure replaces the "offset" fields used earlier in the data collection": see discussion in Section 13.2

The categories for state involvement are the following:

State Perpetrator: Reports explicitly identify the perpetrators as employees of a state agency—i.e., members of the armed forces, the police, other official security forces, or any other government agency—for the state in which the event occurred.

Non-State, Internal, State Sanction: Attackers are identified as members of a non-state organization or group with purported state support or state sanction.

Non-State, Internal, No State Sanction: Attackers are identified as members of a non-state organization or group, and there is no allegation of state support or sanction for their actions.

Transnational: State Attackers are identified as members of a state group from across an international border.

Transnational: Non-state Attackers are identified as members of a non-state group from across an international border. (see comments below)

Multiple Perpetrators (State)

Multiple Perpetrators (State And Non-State)

Multiple Perpetrators (Non-State)

Unknown/Unclear/Other: This category includes cases of "unidentified gunmen", suicide bombers, car bombers and other unidentified explosives where no groups take responsibility.

8.0.1 Ambiguity: Transnational non-state actors

With the emergence of explicitly trans-national Islamic militant groups such as al-Qaeda and ISIS, the distinction between the "Non-State, Internal, No State Sanction" and "Transnational: Non-state" actors has become quite ambiguous, and is becoming even more problematic as some groups which are probably almost entirely from inside the state where the incident occurs "declare allegiance" to ISIS in particular. At present these are generally categorized as "Non-State, Internal, No State Sanction", but standard names are used so this could be changed in an analysis. The fluidity of the Afghanistan-Pakistan and northern Nigerian borders present similar issues for the Taliban and Boko Haram, though the latter is assumed to be based in Nigeria even when it is attacking outside of Nigeria. East African militia groups such as the Allied Democratic Forces, which began in Uganda but have been operating in the Nord Kivu province of the DRC for about two decades are another ambiguous case. Which is to say this category may not be particularly useful and these situations need to be resolved based on the group.

9 Perpetrator identity variables

All of these fields contain descriptive text from the reports of the event that correspond to the category. For example, if the perpetrators are ideological, generally the name of a group will be given, though a generic such as "leftist rebels" is also acceptable.

Military: Attackers are identified as uniformed members of an official state military force. Keywords include "armed forces," "soldiers," "security forces."

Police: Attackers are identified as law enforcement officers.

State Other: Attackers are identified as state agents, but there is not sufficient information to characterize them as either military or police, or they are identified as employees of a different state agency, typically a secret police unit.

Non-State Ideological: Attackers are not affiliated with the state and are identified in terms of their membership in an ideological organization or by their political or ideological views.This category includes any named militant group such as the Taliban, Boko Haram, and ISIS.

Non-State Ethnic: Attackers are identified in terms of their ethnicity, race, or nationality.

Non-State Religious: Attackers are identified in terms of their religion.

Non-State Criminal: Attackers are identified as members or associates of a criminal syndicate, drug cartel, mafia, etc. Criminal activity includes production and trade of illegal drugs, arms smuggling, kidnapping and extortion, and prostitution when this has a coercive element. If the group is clearly involved in both criminal and legal activity (for example a warlord who controls both diamonds and illegal drugs), use both the private and criminal categories. Per the earlier discussion, criminal perpetrators are coded only when the group is exercising sufficient power (e.g. challenging state control) as to be considered "political" at least in some sense.

Non-State Private: Attackers are identified as part of a private army or security service reportedly working on behalf of a warlord, landowner, private corporation, or association of any such individuals or interests (e.g., a cartel of landowners).

Unknown/Unclear/Other: Situations where the identity of the actors is not known.

In the absence of claims of responsibility, the perpetrator identification is sometimes unclear, though local authorities or the reporters attribute it to a group, typically using the word "suspected", e.g. "suspected Taliban" or "suspected Fulani herders." In these instances, the group name is usually entered followed by a question mark ("?") though this coding convention has not been used completely consistently. If there is only one group in the region using the tactics in the incident—for example Boko Haram village raids in the 2014-2015 period, which were quite distinctive—it is attributed to them. In the cases of car bombings and suicide bombings where no group accepts responsibility, there is usually a high prior probability on the likely perpetrator, but generally these are coded as "Unknown": this could be changed on a regional (and temporal) basis in any later analysis.

10 Victims

10.1 Victim Noncombatant Asserted

Noncombatant Status Asserted: At least one report of a given event identifies some or all fatalities as noncombatant civilians, usually by labelling the victims in terms that are mostly or exclusively applied to noncombatants (e.g., "civilians", "people", "villagers," "passengers," "men, women, and children").

Noncombatant Status Not Asserted: Reports are inconsistent as to whether all of the victims were noncombatants: this is generally used in situation where the report indicates that both

combatants and noncombatants were killed, but does not break out the casualties in each category, protests where some protestors seem to have been armed, and the "clash" category discussed in Section 11.1.

10.2 Victim Noncombatant Status Contested

Noncombatant Status Contested: This is generally used when officials say that the victims (typically protesters) were armed and this is disputed by surviving representatives of the victims or other observers. It is also used for the "clash" category discussed in Section 11.1

Noncombatant Status Not Contested: No reports of a given event contain an assertion that the victims were combatants.

11 Victim identity variables

These fields are used to record information about the identity of the victims of an atrocity or atrocities. As with the perpetrator identities, these fields contain text from the report indicating the identity.

Political: Victims are identified as members a particular political party, political organization, social movement organization, issue-oriented non-governmental organization, or in more general terms as sharing certain political beliefs or sympathizing with particular political or paramilitary groups.

Ethnic/National/Citizenship: Victims are identified as members of a particular ethnic, racial, or national group, a particular tribe or clan, or citizens of particular state.

Religious: Victims are identified as members of a particular religion, sect, or spiritual movement.

Socio-Economic: Victims are identified as members of a particular socio-economic class or occupational group (e.g., peasants, landowners, squatters, workers). This is also when the individuals were killed in the context of an attack on property, such as an oil refinery.

Unarmed Combatants: Victims are identified as individuals who under other circumstances would have been combatants, but were unarmed and unable to defend themselves at the time they were killed (e.g. unarmed members of guerrilla groups; off-duty police or security personnel).

Random/Unknown/Unclear/Other: This field should be used for instances of terrorist attacks where the victims appear to have been killed at random; where victims are described as coming "from all walks of life" or similar terms; where victims are described in terms that do not relate to any of the above categories (e.g., "women and children"); or where no description is given.

Coders should record all categories explicitly mentioned in any reports of a given event, but coders should not record any categories not explicitly mentioned in those reports, even if those additional categories are widely known. For example, if a report identifies victims as "Kosovar Muslims," a coder would record both a religious and an ethnic/national identity; but if a report simply identifies the victims as "Kosovars," a coder would only record an ethnic/national identity. Note also that these fields should not be used to capture information about individual victims if that information

is not apparently intended as representative of a larger set of the victims (e.g., details provided in an article that describes the different interests and occupations of individual victims). Finally, coders should record any information that led her to conclude something about the identity of the victims (e.g., "tribal violence" or "sectarian killings") in the `Description` field.

Casualties

These fields are used to record information about the number of deaths and injuries associated with a particular event. For events described in multiple articles, use the highest available figure or estimate, unless a more recent article offers a figure or estimate that is appears to be based on more information than the previously reported figures, in which case the more recent figure should be used instead.

The categories for death and injury scales are as follows:

0: This applies only to the `injurscale` field, typically in targeted assassinations, where it is clear no one was injured except for the individuals killed.

1-4:

5-24: Keywords include "several", "a few", "a number of"

25-49:

50-99: Keywords include "tens", "dozens", "scores".

100-999: . Keywords include "hundreds".

1,000-9,999: Keywords include "thousands".

10,000+: Keywords include "tens of thousands" [this category is no longer used as it only applied to extended "campaigns"].

No information: This applies only to the `injurscale` field and indicates that no information on injuries was reported.

On deaths, when a phrase such as "up to 30," "about 30,", "at least 30,"—probably the single most common construction—or "more than 30" is used, base the category on the reported number ("30" in this example): the point is that some explicit number of required on the `Death` field. In the `Injuries` field, it is okay to use terms such as "several", "many", "dozens" and so forth, then set the `injurscale` to the based approximation based on the apparent magnitude of the attack.

Death Ambiguity

The categories for the ambiguity variable are as follows:

Single Number Reported: Indicates that the report or reports have been resolved to a single number, even though multiple numbers may have been reported in incomplete reports.

Range Reported: Includes cases where only a numeric range is reported: this is very rare.

Multiple Numbers/Ranges Reported: Indicates that the report or reports provided different death counts or estimates in circumstances where one would expect only a single number.

No Numbers/Ranges Reported: Indicates that the relevant report or reports provided no indication of the number of deaths. This is no longer used.

Death Contested

For the reasons discussed in Section 13.6 this is no longer systematically coded, though if are reasons to suspect the reported number these can be described in the `Description` or `Comments` fields.

Mode

The categories for `Mode` variable are as follows:

Attack/Massacre: Single or multiple perpetrators engage in planned attack: this is by far the most common category.

Riot/Pogrom: Group of perpetrators engages in apparently disorganized or rudimentarily organized and relatively spontaneous attacks, typically in the context of ethnic violence.

Protest: Deaths occur in the context of an organized protest, usually against the government. In such circumstances, police almost invariably claim that protesters were armed so the "noncombatant" criterion is less strict. Added December-2013.

Clash: Ethnic violence where both sides are armed: see discussion below. Added December-2013.

Unclear/Other: Rarely used

11.1 The ambiguity of the "clash" category

This category, introduced in 2013, has proven problematic and may still be subject to change. "Clash" is consistently applied to a situation where both sides are armed, both sides are non-state actors and there are casualties on both sides. In almost all cases, these incidents are associated with a "tribe", "clan" or some other ethnic designation; in a few instances they are armed thugs associated with political parties where that sort of organization is common. In these circumstances, the `Victim` fields are left blank and both sides are considered "perpetrators."

So given that the victims clearly *aren't* non-combatants, why are they in this data set at all? The motivation here was to make sure we were getting ethnic violence—or in the case of farmer-herder conflicts, violence possibly linked to climate change—that could subsequently escalate, that is, monitoring and early warning. Which is working to a degree but still has the following issues

- There are some "repeat offenders" who one finds are doing this sort of thing all of the time—this would include certain sub-tribes in Sudan, "cults" in Nigeria, and drug gangs in Central America (which are only marginally "political" in any case)—and there is little or no evidence that it could escalate. In fact to the contrary, there is usually considerable concern by local

authorities that these don't escalate: we're basically dealing with Capulets and Montagues here, not Hutus and Tutsis.

- In some areas, notably Yemen, it is difficult to tell whether at least one of the groups is essentially acting as a paramilitary on behalf of the government, so this is simply unconventional state-led warfare rather than an atrocity.

So, this is still a work-in-progress, and could be dropped or further refined in the future.

Weapons

In ascending order of lethality, the categories for the weapons variable are as follows. If multiple types of weapons are used, record the most lethal.

Primitive Weapons: Perpetrators directly inflict violence on victims at close quarters using primitive weapons such as machetes, spears, clubs, or knives.

Firearms: Perpetrators directly inflict violence on victims at close quarters using small firearms, such as pistols, rifles, and light automatic weapons.[14]

Explosives: Perpetrators use an explosive device that is expected to detonate without killing the perpetrators; this includes grenades.

Suicide Bombing: Perpetrators detonate an explosive device that also takes their lives, apparently by design.

Heavy Weapons: Perpetrators use crew-served weapons such as heavy machine guns, artillery, tanks, or aircraft.

WMD: Perpetrators use chemical, biological, or radiological weapon(s) to carry out a mass killing.

Unclear/Other: Use the `Description` field to record any potentially relevant keywords: this is typically used for deaths caused by fires, notably in prison riots, which may or may not have been deliberately set. This also applies to cases where a vehicle without explosives is used as a weapon.

Perpetrator Intent

For an event to qualify as an atrocity according to the conceptual framework underpinning this project, perpetrators must have intended to kill noncombatant civilians. As noted earlier, however, explicit intent is often difficult to establish, and we generally aim to finesse this problem by assuming intent unless otherwise stated in the journalistic accounts.

Nevertheless, we expect many reports will contain information about perpetrator intent, and we wish to take advantage of this information where it is available. To do so, we have created a variable that provides coders an opportunity to record whether or not intent was asserted; if it was, whether that assertion was disputed. This variable allows us to narrow or broaden the sample of events

[14] We're increasingly getting reports of incidents involving firearms that are described as "suicide attacks," presumably because the perpetrators made no effort to escape. At present, these are just going into the "*Firearms*" category.

used in any analysis, depending on how conservative we want to be in our approach to the question of what constitutes an atrocity.

Note that in all instances, the statements in question may be made by the journalist or by a source the journalist quotes or cites.

Intent Apparent But Not Stated: This is the most common category: The description of event makes evident that perpetrators intended to kill victims in question, but intent is not specifically stated in the report(s). This would apply if actions were undertaken which a reasonable person would assume were intended to cause deaths—for example, firing live ammunition directly into a crowd or setting an explosive device—but the report does not explicitly state that there was intent.

Intent Asserted And Not Disputed: This is used when there is an explicit claim of responsibility for the attack; it is pretty much the same as the "*Perpetrator Approves*" category.

Noncombatants Not Intentionally Targeted: Description of event states that the action by the perpetrators was deliberate, but that the intent was not to kill noncombatants. This will generally be used when the perpetrators were aiming at something else, or when the perpetrators were not aware that the victims were noncombatants, for example attacking a house where they thought there was a sniper. For such events, coders should provide some information about the apparently intended target in the `Description` field.

Intent Asserted/Conflicting Accounts: A single report or multiple reports offer competing accounts of whether or not the perpetrators intended to kill the victims in question. For example, some reports might say that fire was directed into the crowd while others stated that the intent was to fire warning shots over the crowd.

No Information/Unclear/Other: Rarely used: almost always one of the other categories is applicable.

Additional comments on the extent of the specificity/ambiguity of the report can be recorded in the `Comments` field.

A second variable is used to record expressions of approval, regret or apology made by the perpetrators or their leaders or commanders. As usual, clarifying information should be recorded in the `Comments` field.

No Information/Unclear/Other: No expressions of regret, apology or approval are reported within a month of the event's occurrence: this is by far the most common category.

Perpetrator Approves: Reports indicate that one or more organizations has claimed responsibility and approves of the killings; these statements are made by the perpetrators or their organization within one month of the event's occurrence.

Perpetrator Regrets: Reports include expressions of regret or apology made by the perpetrators or their organization within one month of the event's occurrence.

12 Collateral Damage:

This variable is now used to record situations where there was deliberate damage to property in addition to that which is an inevitable consequence of the mode of violence, which is to say that large

car bombs, suicide bombings using vehicles and heavy weapons can be assumed to *always* result in property damage in addition to the deaths. Deliberate additional damage typically involves burning homes, vehicles and businesses in conjunction with a raid, and/or killing livestock and destroying crops. See additional comments on this in Section 13.3.

Collateral Damage: Substantial property damage was mentioned in the article

No Collateral Damage: There was no additional property damage beyond that expected given the mode of the attack.

Unclear/Other: Article does not mention property damage beyond that expected given the mode of the attack.

Related Tactics

See Section 13.4: Only `Assassinations` is now used consistently

Description

This field should contain a short description (typically one or two sentences; frequently the lede sentence of one of the reports) describing the event in narrative English. The description should resemble the lede for a newspaper article, describing the who, what, where, when, and why of the event in question, to the extent possible from the available information. The description should indicate whether the event's status as an atrocity is contested and should provide information about any wider conflict to which the event is apparently linked. Include any keywords that you think might be useful to an analyst searching the database. Information about the credibility of the source reporting the event, and any uncertainty that the coder had in coding various data fields should be recorded in the `Comments` field rather than in the `Description`.

Primary Source and Secondary Source

We use three different types of sources; these are differentiated by the `Source Type` field. Record all of the sources that reported the incident.

International: These are one or more of the international news sources that are used to define the scope of the data collection.

IGO/NGO: These are intended to be reports from major IGOs such as UNHCR, UNICEF, and the International Committee of the Red Cross/Red Crescent, and major NGOs such as Amnesty International, Human Rights Watch, and Doctors without Borders. "Major" means that the group has been in existence for some time and works in multiple areas around the world. Assume that any UN-affiliated IGO is "major." If there is a question about the status of an NGO, assume it is local, even if it works in more than one country.

Local: This is everything else, and includes regional news agencies (e.g. Xinhua, Al-Jazeera), local newspapers, radio and television services, and news agencies of government and militant groups.

`Primary and Secondary Sources`

The primary source is the source of the story that the information has been coded from. A secondary source occurs when the primary source is basing the story on a report from another source, rather than doing original reporting: for example Reuters quoting the NGO Doctors without Borders or quoting a local newspaper, as well as the extensive secondary reports in the BBC and All-Africa. If the story consists of both original reporting and quotes from other sources, then code it as being only a primary source story, but indicate the other sources in the `Comments` field.

`Contesting Sources`

No longer coded: see appendix.

`Citation`

This is use the 25-character Factiva identification number for each of the articles used to code the story, separated by commas. Any additional sources used for the coding should go into the `Comments` field.

Comments and Coder

12.1 Comments

This is a text field that can be used to record any comments relevant to the sources and coding. Most commonly it is used to provide additional information on the geolocation or to provide details about contested or unclear information. Generally `Comments` are used to assess the quality and reliability of the reports, whereas the `Description` is used to provide information about the event itself.

12.2 Coder

Initials of the coder.

13 Assorted random comments and caveats

1. There is now a sizable literature on the difficulties of estimating casualties in conflicts, and anyone using this data is strongly urged to consult that. Though much of this is common sense: we have much better data in areas where violence is relatively infrequent and which are easily accessible to an open, unthreatened and well-funded English-language press than we do for areas lacking some or all of those characteristics. None of these data sets provide a "god's eye view": we're just doing the best we can as constrained by the source reports and the resources for coding these.

2. Three atrocity patterns where I think our data are particularly weak:

Full scale wars: These have multiple issues, including the sheer volume of the reports, the difficulty of access and "media fatigue" where the details of the war are no longer seen as newsworthy. Specialized sources are almost certainly more useful for these.

Targeting of journalists: This has almost certainly affected the reporting on drug violence in Mexico, Honduras and El Salvador, and probably some parts of Pakistan (e.g. Karachi).

State sanctioned police killing campaigns: These tend to be very diffuse in both time and geography, and the international press generally only reports aggregate numbers without details. The extra-judicial killings of alleged drug users and dealers in the Philippines following the election of Rodrigo Duterte, which reportedly involved over 2,000 deaths in a two-month period, did not generate a single report of a discrete incident involving more than five killings. Investigative reports by human rights groups are probably the better source here.

3. Routine election violence—as distinct from the large-scale violence seen, for example, in Kenya in December 2007 and January 2008—tends to be very dispersed and only individual incidents above the 5-death threshold are reported: in other words we don't code all of the violence in conjunction with an election as a "campaign." When candidates and party officials are the victims, these are recorded as targeted killings.

4. Land mines that were not deliberately triggered and children playing with previously unexploded munitions (UXO) are not coded. The former is definitely is a source of ambiguity since it is not always clear whether an explosion was a deliberately-triggered IED: if there is uncertainty, it is coded.

5. Darwin Award candidates do not make it into the data. This would include cases where a qat-vendor accidentally pulls the pin on a grenade in their pocket while reaching for their car keys (Reuters, 26 Sep 2012), suicide bombing instructors accidentally demonstrating with a live vest (most emailed, everywhere, 10 Feb 2014), the almost monthly cases of individuals testing the efficacy of recently-purchased bullet-deflecting magic charms by having someone shoot at them,[15] and some members of the Taliban using a hacksaw on a mortar round to make an IED and killing their entire household. (Mar 2014).

6. When a militant group establishes territorial control over an area, reports often drop dramatically. This was certainly true for ISIS, which also had an extensive media operation of its own, and also during the short period when Boko Haram took control of cities rather than making hit-and-run attacks on villages.

7. We have *generally* copied information on perpetrator and victim identities from the articles themselves rather than trying to standardize the identification. This leads to an assortment of transliteration inconsistencies, such as "Muslim" and "Moslem" or "Sh'ia", "Shia" and "Shiite." Groups also go by various names, notoriously the entity variously known as Islamic State, IS, ISIS, ISIL and, lately, Daesh.[16]

However... the twin technological imperatives of Excel (pre-2013)[17] and the Chrome browser (post-2013) retain copies of recently-used entries in various fields, so, for example, almost all of the "ISIS" entries use "ISIS" or "Islamic State" because these can be filled in with auto-completion[18] even if the report, say, used "ISIL" or "Daesh." In the on-going effort to provide a cleaned version of the

[15] Curiously, these tests always seem to be done on the purchaser, not the vendor.

[16] See `https://www.freewordcentre.com/explore/daesh-isis-media-alice-guthrie` for an excellent exposition on the last.

[17] From *Science* 2 Sept 2016: 20%—Fraction of genetics papers in top scientific journals that contain errors in gene names due to autoformatting in Microsoft Excel. For example: Septin-2, often called SEPT2, is "corrected" to the date 2 September (*Genome Biology*).

[18] Or through the template files which are used, for example, for Iraq, Somalia and Nigeria.

data, persistent groups will be resolved to their TORG number and a standard name,[19] but we're not there yet.

8. During the first Palestinian *intifada* I was involved with an East Jerusalem-based NGO that documented fatalities in that conflict, and saw first-hand the difficulties in generating this type of data even in an area which was geographically compact, had excellent communications infrastructure, and where journalists and NGO workers were only rarely targeted. I can only begin to imagine the amount of time, effort and personal risk that has gone into securing the information that has gone into these reports, and the courage of those who will talk to reporters: the global community is deeply indebted to those who are working to see that these deaths do not go unnoticed.

[19] `http://www.albany.edu/pvc/data.shtml`

Appendix: Summary of all fields in current data format

The data are provided in an Excel spreadsheet. The following types of entries are used:

Table 1: Variable value types

Text	Any text can be entered
Date	Date in the form dd-mm-yyyy, e.g. 05-12-1997 for December 5, 1997
Number	Numeric value
Category	Value is chosen from a fixed of choices
Boolean	True/False

Table 2: Field vames and types

Event Type	select from categories
Campaign Identifier	[no longer coded]
Event Reporting	select from categories
Start Day	dd
Start Month	mm
Start Year	yyyy
End Day	dd
End Month	mm
End Year	yyyy
Country	select from ISO-3166-alpha-3 codes
Region	text
District	text
Locality	text
Latitude:	
Degrees	dd
Minutes	mm
Seconds	ss
Direction	N or S
Longitude:	
Degrees	ddd
Minutes	mm
Seconds	dss
Direction	E or W
Offset Distance	[no longer coded]
Offset Direction	[no longer coded]
Perp State Role	select from categories
Perp State Military	text
Perp State Police	text
Perp State Other	text
Perp Non-State Ideological	text
Perp Non-State Ethnic	text
Perp Non-State Religious	text
Perp Non-State Criminal	text
Perp Non-State Private	text

Perp Unknown/Unclear/Other	text
Victim Noncombatant Status Asserted	select from categories
Victim Noncombatant Status Contested	select from categories
Victim Identity Political	text
Victim Identity Ethnic	text
Victim Identity Religion	text
Victim Identity Socio-Economic	text
Victim Identity Unarmed combatant	text
Victim Identity Random/Unclear/Other	text
Deaths Number	record number if given
Deaths Scale	select from categories
Injured Number	number or adjective
Injured Scale	select from categories
Organization of Violence	select category
Weapons	select category
Deaths Ambiguity	select from categories [only rarely coded: see notes]
Deaths Contested	select from categories [no longer coded]
Intent	select category
Regrets	select category
Collateral Damage	select category
Description	text
Link	[no longer coded]
Primary Source Type	[select from list
Primary Source	text
Secondary Source Type	select from list
Secondary Source	text
Contesting Source Type	[no longer coded]
Contesting Source	[no longer coded]
Citation	text
Comments	text
Coder	text

Appendix: Syria

When the coding transitioned to Parus Analytics in 2013, we shifted from coding casualties in Syria using the international sources to using a specialized NGO source, `http://syrianshuhada.com/`. The reason for this was that because the conflict was so intense, the reports in the international sources were—except for those involving very high casualty incidents—essentially random and largely a function of where reporters happened to be on the ground, rather than providing a reasonably consistent overall record. At the time we made this decision, there were four NGOs providing information on civilian deaths just in Syria, and after reading several comparative assessments of these, we chose `syrianshuhada.com` because it seemed to be the most consistent and used fairly conservative inclusion criteria.

By early 2016, however, updates to that site were increasingly delayed, hampering the ability to provide monthly updates, so we transitioned to a new source:

`http://www.vdc-sy.info/index.php/en/`

which is used, for example, by *The Economist* and was one of the two data sources integrated into the `syrianshuhada.com` data, as well as being very web-scraper-friendly, allowing easy extraction of the case-by-case data. However, the definitions (and presumably methodology) of the two sources are somewhat different and, if I'm reading correctly from their earlier summary reports (see for example the May-2015 report at `http://www.vdc-sy.info/index.php/en/reports/1433810787`, these don't include fatalities due to the anti-regime forces except for ISIS and "the coalition forces against ISIS". However, a test with the data for July-2015 to April-2016 showed the daily counts correlate quite well ($r = 0.800$). The VDC counts are about 60% lower and this lower number is probably a combination of

- the exclusion of deaths caused by rebels
- inclusion of rebel fighter deaths in `syrianshuhada.com` : these are about 25% of the cases and were included in the earlier collections because we couldn't get daily totals that distinguished civilian and rebel deaths.
- other differences in the information-collecting methods: see `http://www.vdc-sy.info/index.php/en/about` for more details on the VDC methodology.

Note that both of the sources are explicitly anti-regime and `syrianshuhada.com` notes on their home page that their data is a combination of the VDC data and data from the Damascus Center for Human Rights Studies. As with everything in this data set, the use of these sources does not imply any sort of endorsement by the US government.

In addition to providing information on the civilian/military status of every victim, the VDC data also provide geographical information at the province level, and causes of death (shelling, firearms, torture, etc). These codings are preserved in the new data so we now have geographically disaggregated data. Provincial coordinates are the geographical centroids given in `geonames.com`; the "Other" and "Unknown" locations have been assigned to points vaguely near a population sort-of centroid. The 5-death threshold has not been applied at the individual region level (it is always exceeded on the country level), and all entries are labelled as being part of a "campaign" even if there was only a single incident for a given region/day.

In addition to making the substitution of the source, the default values on some of the fields have been updated to reflect the new data:

- `victimrand` is "civilians" since we now only are counting victims explicitly identified as civilian
- The `perprole` category is "*Multiple Perpetrators (State and Non-State)*"
- Four weapons categories—"*Heavy Weapons*"','"*Explosion*"','"*Firearms*" and "*Other*"—are now distinguished, in this priority (that is, the event is tagged with the most severe weapon used)
- `injurscale` is set to "*No information*" and `deathambig` is set to "*Single Number Reported*".

Appendix: Variables no longer collected

First, a bit of background as to why we even have this appendix. Standard practice for social science data sets, of course, involves working out a codebook, refining it through multiple iterations in pilot studies, then after finalizing that codebook, staying as closely as possible to the original definitions for the life of the project. But that's obviously not what we've done here: why?

As noted in the introduction, the entire purpose of the data set has been for applied monitoring and early warning rather than classical hypothesis testing. Consequently from the beginning, the possibility was left open that the coding could change, and the contract authorizing the data collection actually specifies that PITF be notified on a monthly basis of any such changes. Major changes are made less frequently than that, but over the roughly twelve years of the project, several have accumulated. These are generally due to six factors:

- The original scheme was still significantly influenced by the focus of the "State Failures Project"—the predecessor to PITF—on potential genocide (and more generally, the models of atrocities seen during the 1990s in the violence in Liberia and Sierra Leone, Rwanda, and the former Yugoslavia), which because of its scale requires resources similar to those of a state and nearly complete state failure. In the 21st century, violence against noncombatants is primarily perpetrated by non-state actors and usually in circumstances where a functioning state exists and is fighting those actors.
- The dominant mode of killing has shifted to the use of explosives and various forms of suicide attacks which preclude a number of tactics seen in earlier episodes of violence against noncombatants such as hostage-taking and sexual violence.
- A small number of characteristics which we expected to be relatively common in news reports—particularly the credible contesting of whether an atrocity occurred—turned out to be very rare. Again, this is partly a function of the shift from state to non-state perpetrators: non-state actors tend not to have the resources (or don't care) to monitor and contest media reports, and in recent years, many actually prefer to exaggerate their violence against noncombatants rather than deny it.
- After coding several thousand incidents from several tens of thousands of reports, it became clear that some of the things we had intended to code were largely a function of the type of violence, for example the multiple casualty counts and implied property damage in car bombings, rather than specific to the reports.
- The original coding scheme envisioned a complex set of incident/campaign linkages which proved to be impractical because of the impossibility of systematically determining whether various incidents were connected.
- There were a number of changes implemented in 2013 when the coding shifted from a coding group working in Kansas to Parus Analytics: These reflected lessons learned and new resources than were not available when the project started in the mid-2000s.

In a world of unlimited time and resources, we would go back and recode everything to a single standard, but... shock, shock... we've had neither, and for the purposes for which the data have been funded, these changes have been generally unproblematic: Most involved relatively marginal aspects of the data and the core "who killed whom, how, where and when" variables have not been affected. So... well, it is what it is.

That said, as time permits, we are gradually working on a "cleaned" version of the entire data set which to a large extent will apply the criteria in this version of the codebook to the entire data. Specifically, the revisions focus on

- Each incident will be assigned a unique identifying key
- Geolocating events which were previously coded only to the province level—often with centroids—to localities where possible
- Eliminating large-scale aggregated events and any other events that cannot be resolved to a locality and week
- Standardizing all categorical fields, and in particular removing typos that occurred when the data were coded using an Excel spreadsheet rather than the current web-based coding form
- Standardizing dates: the original codebook allowed blank fields when dates were unclear
- Breaking out non-numerical death and injury estimates into separate fields so these totals can be read as integers by statistical programs
- Standardizing militant group names and adding TORG codes for these
- Removing unused fields
- Guaranteeing that the data comply with the .csv "standard" as well as providing an alternative version in the JSON format.
- "Targeted killing" will become a `mode`

The remaining subsections discuss the major changes since the original version of the codebook and provide some background as to why these were undertaken.

13.1 Geographical locations

Here we've substantially tightened the criterion, due partly to our ever-increasing ability to geolocate thanks to improvements in various web-based tools, and also increasing use of geolocation in the PITF models. Current geolocation is much more precise than the geolocation in the early data; as noted above, the "cleaned" dataset should have relatively uniform geolocation based on contemporary tools.

13.2 Geographical offsets

The original data had an "Offset" option designed for phrases such as "10 km north of..." Unfortunately, such directional statements were a lot less frequent than variations on "10 km outside of...", where only a distance is provided without the direction. Offsets also require a fair amount of processing—albeit from readily available code—to convert to coordinates, so given the imprecision, it seemed better to just provide an approximate coordinate based on `maps.google.com`. When approximate coordinates are used, this is always documented in the `Comments` field.

13.3 Collateral Damage:

This variable has had two problems. First, as originally described in the codebook, it was equivalent to the "Noncombatants Not Intentionally Targeted" category in `Perpetrator Intent`, with the assumption that we would find a reasonable number of cases where casualties occurred in the context of large military operations. In fact, these are not particularly common: this may in part be due to a reporting bias—as one gets fewer details out of areas where there is intense conflict (this has been particularly problematic in Yemen and may be an issue in Afghanistan)—but also is affected by the shift from state to non-state perpetrators. It also frequently turned out to be difficult to distinguish whether or not targeting was deliberate given the tendency in contemporary warfare for terrorizing noncombatants either to deprive local militant groups of support, to drive them out of a territory, or to kill them as part of an "ethnic cleansing" campaign.

As a consequence, the `Collateral damage` field drifted to being used to code extensive *property* damage, but even here it has generally been used only to record situations where there was damage in addition to that which is an inevitable consequence of the mode of violence, which is to say that large car bombs, suicide bombings using vehicles and heavy weapons can be assumed to *always* result in property damage in addition to the deaths, even that is not explicitly coded in this field.

13.4 Other tactics

In the original codebook, we had planned to code the following:

- Interference with delivery of food or aid
- Scorched earth tactics
- Use of human shields
- Rape and other forms of sexual violence
- Assassinations: this has been expanded and coded systematically in the "targeted killings" category
- Detentions
- Sieges and closures
- Kidnapping
- Disappearance

This didn't work out due to several factors

- Denial of food and aid, detentions, sieges and closures and disappearance tend to be state-level *strategies*, rather than tactics associated with specific incidents by non-state actors
- We've found almost no instances of human shields
- Sexual violence tends to be less commonly associated with atrocities incidents—it is obviously still common in conflict zones generally—now that the dominant mode is killing through car and suicide bombings. Also there are better data sets for sexual violence, particularly `http://www.sexualviolencedata.org/`.

- Kidnappings are difficult to sort out from the criminal variety, as well as being less common as the dominant mode of violence has shifted

While not systematically coded, these categories are still reported in the `Description` field when they seem relevant to whether the incident signals an escalation in violence.

13.5 Link

This was originally designed to connect incidents that appeared to be related but, like "campaigns" more generally, "related" required us to know the strategic plans of the perpetrators (who are often anonymous), and we don't. Information on the perpetrator, victim and location is usually about as close as we can get to this.[20]

13.6 Contesting Sources

In the original version of the codebook, there were a number of fields which were intended to code the degree to which a report was uncertain or contested. These proved impractical to implement consistently and therefore we stopped trying to code them. That impracticality came from two major sources

1. The ambiguity of casualty totals is largely a function of the type of attack

And in particular, the casualties in a mass attack—except those in developed areas (e.g. Paris, Istanbul)—are *always* uncertain, both due to the difficulty of actually counting the number of victims (without going into the sorts of details that make coding these things less than pleasant, consider the effects of large explosions. . .) and the fact that the international news sources generally do not provide any followup on how many wounded have died. A different set of problems occurs in attacks on remote villages, whether it can be unclear whether missing individuals escaped and are in hiding or otherwise displaced, were kidnapped (particularly young women), or were dead and in the burned remains of destroyed buildings. In targeted killings, in contrast, the number of dead and injured is usually unambiguous.

2. The likelihood that someone will "contest" a source is largely institutional rather than a function of the event itself

Perhaps surprisingly, in most places reports are not contested at all, apparently because the perpetrators and the authorities do not care what is being reported. When they are contested, the alternative reports are generally so ludicrously consistent (e.g. authorities always reporting much smaller death totals or perpetrators reporting higher ones, albeit typically of combatants) in comparison with the eyewitness testimony that there seems little point in reporting these.

The rise of "cell phone journalism" probably has also discouraged state authorities from contesting reports in areas where there is an active press: denying events simply encourages more reporters to try to figure out what actually happened, and they can do this at a relatively low cost. So authorities who once would have disputed a report now simply keep quiet and hope it won't be noticed.

[20]There is also a cryptic note in the original codebook that the "Link" field was supposed to provide a unique key that could be used in a relational database: that was never implemented but will be provided in the forthcoming cleaned version of the data.

A comment by an eyewitness that probably applies to most of these disputed casualty reports:

> An eyewitness, Sani Skandi, who spoke to our correspondent on the clash said over 15 people died. "I counted over 15 bodies. But if [the authorities] told you only seven people were killed, so be it," he said.

In a small number cases, there will be a genuine ambiguity of casualty totals based on eyewitness reports from individuals who have access to information and no apparent institutional incentive to distort the figures (e.g. when multiple reports are received from a remote area): these are noted in the `Comments` field.