**“The Trials and Tribulations of Collecting Data on Violent Non-state Actors.”**

**Corina Simonelli**

**Victor Asal**

**VERY PRELMINARY DRAFT – COMMENTS WELCOME**

Incident Level vs. Group Level

As with any data collection effort, obtaining and ensuring the accuracy of organizational and network information on violent non-state actors presents a number of challenges. These challenges appear at all stages of the data collection, from establishing a codebook, to identifying and accessing reliable sources, to interpreting the sources accurately and with consistency, to correctly converting the information to data and finally during quality control stages. This paper examines some of the problems that have been identified by the Big Allied and Dangerous (BAAD) team and offers solutions to mitigate their impact for similar collection efforts.

Inclusion criteria and codebook. The first stage is the creation of a codebook and determining which groups will fall within the scope of the dataset. The BAAD inclusion criteria are based on a group’s occurrence in event-level datasets such as the Global Terrorism Database (GTD)[[1]](#footnote-1) at the National Consortium for the Study of Terrorism and Responses to Terrorism (START) and Uppsala Conflict Data Program (UCDP)[[2]](#footnote-2). Gathering from these sources resulted in an initial perpetrator list contained over 2,700 unique groups.[[3]](#footnote-3) This included a whole gambit of groups, ranging from international recognized terrorist organizations and small ethnic groups, to obscure Greek anarchist groups and Mexican cartels. Without a prevailing definition of the term terrorist within the academic field, much less a typology of the defining characteristics of a terrorist group, inclusion criterion for a group-level dataset becomes a significant challenge. Firstly, by relying on other datasets, our group lists are at the liberty of differing definitions, and potential mistakes within other projects in identifying groups. For example, the Global Terrorism Database lists several different names for the Islamic State (IS), representing the group’s evolution and the name they operated under at the time they conducted the attack.[[4]](#footnote-4) However, this isn’t necessarily consistent with the UCDP actor dataset, nor represents name changes for BAAD’s coding rules. This is a very common occurrence and requires research or prior knowledge to group Al-Qa’ida in Iraq (AQI), Islamic State of Iraq and the Levant (ISIL), Al-Tawheed Wal-Jihad, Islamic State of Iraq (ISI), ect. as a sole organization rather than several that require coding. While this example is well known in the media, there are several examples where information is less clear. Secondly, common characteristics of small cartels operating in the United States are vastly different from the type of variables needed to encompass the full range of leftist insurgencies in South East Asia.

The first hurdle is easiest to tackle by cleaning and refining the group list before beginning the coding effort. One way to do this would be by creating more specific inclusion criteria to only code a subset of the groups. For example, only code groups of a certain minimum size, focus on groups with a certain ideology or within a region, or limit the project to groups that have committed fatalities. However, whenever restricting along these lines, full scope of the project and its resultant contribution of the field of organizational research is likewise more concise. The BAAD project wanted to capture the characteristics of the full range of groups under the banner of “Violent Non-State Actors”. In order to mitigate the inclusion criteria issue, we devised a Terrorist Organizational List (TORG) that captured information from all our different inclusion datasets. Within this list we were able to condense aliases and name changes either between or within datasets into one group. We also used this list to weed out perpetrators from lists that were generic terms, ethnic groups, or individuals. However, this is just a first step in mitigating the problem. Networking and group variances as a result of large inclusion criteria cause many problems which will be addressed later in this paper.

The second issue, the mass variety in subject matter presents an issue for crafting the codebook. It is difficult to create a codebook that encompasses groups such as Al-Qa’ida or the Revolutionary Armed Forces of Colombia (FARC) and the Beltran Leyva Cartel – Valdez Villareal or Black and Red Anarchist and Anti-Authoritarians Initiative. The initial approach employed at BAAD was to cast a large net, including as much information and as many possible relationship types as possible to accommodate each unique situation. However, this caused inconsistencies within variables, and a mass collection of data that couldn’t readily be useful. This strategy as quickly changed to a more exclusive philosophy. The BAAD project still collects over 40 different organizational variables and 13 networking relationships. However, by not attempting to accommodate every possible group relationship, coding efficiency has increased markedly with no significant loss in content of the data.

Scarcity of information and reliability of sources. Once a baseline codebook has been created, the first step in data collection is finding information from reliable sources. While this is an issue that complicates all coding efforts, it is particularly prevalent for terrorist organizations which may operate to gain media attention for their ideology, but attempt to keep organizational level information covert. Particularly when collecting yearly organizational data, there are several years where there may be no mention of the group’s structure, leadership or funding mechanisms. This results in many missing values. There are many groups that have perpetrated attacks identified in event-level datasets but there is no information regarding their organizational or network variables.

 Once a wide range of sources including Lexis Nexis, Google Scholar, government and international think tank reports, and region-specific policy centers have been scoured with no avail, a group is deemed not code-able. However, determining the criteria for this is difficult. For example, if one article is found describing a leftist ideology for the group but no other information can be found, should that information be entered despite every other variable missing? Moving on to the next source issue, what if that article is from a blog or some less credible source?

Articles describing violent non-state groups often present a bias, either for the group or against it, and this tends to shade the facts in one direction or another. We have to ask ourselves, can the information on this group’s website be trust? If Iran is accusing Israel of funding a group, or visa versa, can we take that as factual evidence of funding even if there may be political incentives? Event-level data might more easily lend itself to specific source type, for example only using news articles. For example, the Global Terrorism Database has a very refined process of sorting through news articles to pull information on an attack including perpetrators, causalities, weapons, location and targets.[[5]](#footnote-5) While these collection projects still encounter issues relating to the credibility of news sources, they are able to limit their scope of sources by focusing solely on news article. However, with the organizational-level there is a full wealth of information that exists beyond these sources. Yet removing these parameters invites in the potential for poor, inaccurate or untrustworthy sources of information; particularly when the information gathering is in the hands of several different undergraduate coders. Allowing coders to conduct research from the full range of web sources available to them presents a number of challenges and needs to be closely monitored. The alternative, limiting the sources they can access, invites the potential for missed information.

It is very important to establish and maintain a list of good versus bad sources. This should include generic descriptions and as well as lists of specific sites. For example, generally, independent news articles, in-depth reports from well respect think tanks, most government publications can be trusted as accurate, although there certainly is an occasional exception. Conversely, personal blogs, opinion pieces in the newspaper, government accusations, and the militant group’s media outlet are not considered reliable enough to use. Another way to counter this issue is to build in source reliability scale options into the coding. Thus the coding still captures some information even if it is presented with the caveat that it might not be trustworthy. In the BAAD network matrix, groups can have a relationship of allies or suspected allies. One criterion for the suspect ally coding would be if a second tier source claims the groups are allies. Within the organizational level coding we have rules for specific variables in how to handle information that comes from a less than top tier source. Often times, if a group is small and unknown, the only ideology information that can be found is from their website and manifesto. Although these would not generally be considered independent and reliable sources, in these cases we code the information. However, if the website also boasted of larger size numbers that could not be corroborated through other sources, this information would not added to the data. This allows for data users with looser source parameters to access information on these unconfirmed relationships.

Intercoder Reliability and Group Variance. When working with a large number of undergraduate coders there are a number of issues regarding training, reliability, flexibility to class schedules, and commitment to the project. However, even once reliable coders have been identified, the biggest challenge is insuring students correctly interpret information in identical fashions. While one student might interpret a group’s ideology as anarchist, another student using the same source could determine the group to be anti-globalization and a third still might code the groups as both. Given each student’s unique perspective, it is important to create a series of rules with the ultimate goal that students would not have to make any decisions. They would simply reference the codebook to see how their specific situation applies to the variables. This is particularly difficult in group-level collection because of the great variance in group behaviors. No two groups are the same and they often don’t fit perfectly into one category. For example, while Al-Qa’ida is organized in a network structure, they operate under a single leader at the top of a hierarchy and maintain a shura council. This could potentially fit within the hierarchal category, network category or governing council category. Further research shows that the true decision making power remained, until 2011, with Osama Bin Laden despite the other leadership mechanisms. However, a group with a similar structure but with more power vested within the shura council would be coded different.

The first step in resolving this issue is to take decision-making power away from the initial coder. If the coder simply codes the group one way without consulting a supervisor, then there is no way to ensure the next coder would assign the same values to a similar situation. Instead, decisions are left to project leaders who can alter the codebook to reflect the coding scheme for that unique situation. Resultantly, the next time a coder comes across that structure; there will be consistency in the way it is coded. Constantly making the codebook more thorough and specific is the best way to mitigate the issue of intercoder reliability. Whenever a coding decision is made, it must be shared with the entire team, not just the coder working on the specific group.

Quality Control. The best quality control mechanism for group-level collection is duel coding. However, it is highly inefficient and often limited resources make it unfeasible. The issue of group variance also poses a challenge for quality control. With group and network coding there are fewer codings that are red flags to a supervisor. With event-level data, a supervisor could easily notice and double check codings that list, for example, a group with one attack in a country across the globe from all their other attacks, or note mismatches between fields. In group-level collection there are very few instances that a supervisor without prior information of the group can look at the data and see to be incorrect. A group will never be leftist and rightist; however any ideology could fit with any structure type which doesn’t reflect on the size, financing, political activity or network behaviors of an organization. Beyond basic data cleanness, the supervisor conducting quality control must either have prior knowledge of an organization, or verify each data point to the source. The latter is an extremely time consuming process that results in a similar resource usage as double coding.

 While the BAAD project is always striving to better this process, improvements in source documentation, student oversight and identifying group-level trends have increased the data accuracy. Students tis each data point to a specific portion of the source document. A supervisor can then verify or negate their interpretation of the source information fairly quickly. Additionally, by identifying trends, it is easier to note and double check any outliers within the dataset. For example if most groups that control territory have a similar structure, it is a good quality control measure to double check any groups that control territory but are coded as a different structure.

Some additional non-state violent organization collection hurdles.

Groups are not finite like events which take place over the course of a few hours or at most a couple of days. There is no natural time frame which completely encompasses each group. Some groups last for years and even decades, while others only perpetrate a few attacks within a single year. Collecting at the year level mitigates this some by allowing organizational changes to be shown over time, but this still does not encompass the full complexity of groups. Many groups may employ a tactic at the beginning of a year and not at the end, presenting coding challenges. Specifically in dealing with network structures there are constant changes. Group’s relationships with one and another are not static and very from year to year or within a few months. As the timeframe becomes smaller the issue of limited sources is exasperated. It is difficult to find a source for every single year listing the financing of a group. This either leads to many missing values or the creation of specific rules on interpolation.

Groups disappear. Determining the demise of a group can be very difficult. Often they simply vanish from the source materials. There is no indication that they have surrendered or been defeated, but there is no further evidence of the group’s existence. In these cases we have to establish when to stop coding. Is the group simply dormant and capable of perpetrating future attacks, or have they truly ceased to operate as an organized group?

Networking relationships. These relationships are particularly challenging as they encompass the full complexity of various groups. One of the biggest issues is determining armed wings, political wings, splinter groups, factions and founding groups. Under the constant strain of limited and unreliable information it is often difficult to sort out the true relationships between groups. This is a hurdle that while we constantly seek to refine the process, we have made significant progress.

1. National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2013). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd [↑](#footnote-ref-1)
2. "UCDP Actor Dataset 2.2-2014, Uppsala Conflict Data Program, www.ucdp.uu.se, Uppsala University" [↑](#footnote-ref-2)
3. “Big Allied and Dangerous (BAAD)” Project on Violent Conflict. http://www.albany.edu/pvc/ [↑](#footnote-ref-3)
4. Global Terrorism Database [↑](#footnote-ref-4)
5. National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2013). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd [↑](#footnote-ref-5)