

Global Terrorism Database: Real-time Data Collection Pilot Evaluation

*Report to the German Federal Foreign Office and the United
Kingdom Foreign, Commonwealth, and Development Office*

February 2023

ABOUT THIS REPORT

The authors of this report are Erin Miller and Brian Wingenroth at the University of Maryland. Questions about this report should be directed to eemiller@umd.edu.

The Global Terrorism Database™ is funded through grants and contracts made to START. This research was funded by the German Federal Foreign Office and the United Kingdom Foreign, Commonwealth, and Development Office. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the views or policies of any funding agencies, the University of Maryland, or START. An earlier version of this report was delivered to project sponsors in June 2022.

ABOUT START

The National Consortium for the Study of Terrorism and Responses to Terrorism (START) is a university-based research, education and training center comprised of an international network of scholars committed to the scientific study of terrorism, responses to terrorism and related phenomena. Led by the University of Maryland, START is a Department of Homeland Security Emeritus Center of Excellence that is supported by multiple federal agencies and departments. START uses state-of-the-art theories, methods and data from the social and behavioral sciences to improve understanding of the origins, dynamics and effects of terrorism; the effectiveness and impacts of counterterrorism and CVE; and other matters of global and national security. For more information, visit www.start.umd.edu or contact START at infostart@umd.edu.

CITATION

Miller, Erin and Brian Wingenroth. "Global Terrorism Database: Real-time Data Collection Pilot Evaluation." College Park, MD: START, 2023.

TABLE OF CONTENTS

Executive Summary	1
Introduction	2
Global Terrorism Database: Workflow Overview	2
Four Stages of Data Collection	2
Data Collection Workflow: Additional Considerations	3
Data Collection Timeline.....	3
Triaging Collaboration.....	4
Triaging Variables.....	5
Real-Time Data Collection Pilot: Project Design	5
Comparison Datasets.....	5
Research Questions	7
Adapting the Triaging Workflow	7
Location.....	7
Perpetrators	7
Targets.....	8
Weapons/Tactics	8
Casualties/Consequences	8
Additional Variables	8
Results	9
Event Records Created, Added, and Deleted	9
Event Records Changed	11
Preliminary Data vs. Full Data.....	13
Process/Efficiency	14
Conclusions	16

Executive Summary

- In 2021, the GTD research team planned and executed a pilot project to evaluate real-time data collection. The team triaged news articles in real time for one month (April 2021) to assess the quality of preliminary data in comparison to data for the same month collected under normal processes, including a significant lag behind real time and multiple layers of review.
- When comparing the data collected at various parts of the triaging process, many attack records (67%) remained unchanged after the initial day they were documented, and most (84%) remained unchanged as of approximately two weeks following the attack (one week after the initial attack record was created). The updates made to the remaining 16 percent of events were often superficial corrections rather than substantive developments.
- When comparing the data collected during the real-time triaging exercise in April 2021 to data collected with the benefit of additional news articles published in May 2021, and additional stages of review by the data collection team, the substantive consistency of the data was very high overall, but varied depending on the type of information:
 - Province/State: 97%
 - Hostages Y/N: 96%
 - Perpetrator Group: 95%
 - Number of People Killed: 91%
 - Number of People Injured: 89%
 - Weapon Type: 84%
 - Target Type: 76%

The variables that were least consistent involved challenges related to comprehensively capturing detailed information for attacks that were more complex as a result of having multiple types of weapons used or multiple types of targets attacked.

- There were several serious logistical considerations that would need to be resolved in order to routinely produce preliminary data, especially in real time. These include issues related to sustainability of the effort, the need for extraordinary collaboration and communication, and obvious inefficiencies in the process. Several of these challenges would be improved upon by performing data collection within a few months of attacks happening, rather than within a few weeks.

Introduction

In 2021, the Global Terrorism Database (GTD) research team planned and executed a pilot project to evaluate real-time data collection for the GTD. The goal of this was to assess two aspects of reliability and validity of real-time data collection: 1) the accuracy of information published about terrorist attacks in the immediate aftermath of the event, and 2) the effect of several stages of review and quality assurance built into the typical GTD workflow.

One of the enduring goals of the GTD team is to reduce the lag in data collection behind real time. However in doing so, it is important to understand the practical implications, including the strengths and limitations of early information published about events. Some aspects of early reports may be relatively accurate, and important to document as quickly as possible for early detection of emerging patterns. Other aspects of early reports may be based on preliminary, unconfirmed accounts that are subject to error. Analyzing the reliability of early information allows the team to provide users with important guidance about its value and prioritize accordingly. Although one of the key strengths of the GTD is the depth of detail it contains about each attack, the results of this pilot project can inform a strategy for producing and using an abbreviated, preliminary version of the data.

An additional outcome of the pilot project was insight into the logistical challenges of real-time data collection; how these challenges impact the overall efficiency of the data collection workflow and introduce unanticipated threats to the accuracy of the data. The following report describes the relevant aspects of the GTD workflow, the design of the pilot project, and the results of the evaluation with respect to both data quality and process.

Global Terrorism Database: Workflow Overview

Four Stages of Data Collection

The GTD collection workflow takes place in a Data Management System (DMS) developed specifically for this purpose. It relies on subscription access to approximately two million news articles published daily around the world and aggregated on commercial platforms, including Lexis Nexis and BBC Monitoring. The general data collection workflow includes the following steps:

1. **Pre-Processing:** A software architect archives source documents published in a particular month and applies Boolean filters, natural language processing, named entity recognition, machine learning, and clustering techniques to prepare the articles for systematic review by subject matter experts.
2. **Triaging:** Subject matter experts review news articles that have been classified as potentially “relevant” to systematically identify individual events that meet the inclusion criteria for the GTD. These researchers create an initial record of the attack and document minimal details about the event in a structured data table.
3. **Coding:** Student interns and part-time research assistants under the supervision of full-time researchers record detailed information about each attack. These personnel are organized into small

“Triaging”
Subject matter experts review news articles that have been classified as potentially “relevant” to systematically identify individual events that meet the inclusion criteria for the GTD.

teams by coding “domain” including: location, perpetrators, targets, weapons/tactics, and casualties/consequences.

4. Final Review: Researchers review the coded records for consistency, add several holistic variables, and remove any records that fail to meet inclusion criteria. The software architect and program manager process files to enforce uniformity of coding conventions and continuity with historical data collection.

These steps take place in an assembly line workflow in monthly batches. For example, while the research team is finalizing data files for events that took place in January, the domain coding teams are conducting the detailed coding of events that took place in February, the researchers are triaging source documents to identify and create initial records for events that took place in March, and the GTD software architect is pre-processing source documents published in April. This workflow typically involves fluctuations in timing between the four stages, depending on the availability of resources and personnel that vary across the academic calendar. As a result, certain calendar months of data collection are typically in various states of completeness as the team pieces information together.

Data Collection Workflow: Additional Considerations

The typical GTD collection process involves several features described here to provide additional context. The team adapted certain aspects of these practices to accommodate the real-time pilot.

Data Collection Timeline

The data collection timeline for the GTD typically lags behind real time, but the length of the lag varies considerably. There is a minimum lag of a few months due to ebbs and flows caused by the academic calendar, but this can be extended due to increases in the number of terrorist attacks that occur or the number of news articles published about the attacks, competing tasking for the research team, lapses in project funding, and turnover in research personnel.

Figure 1. Approximate Timeline: Standard GTD Collection

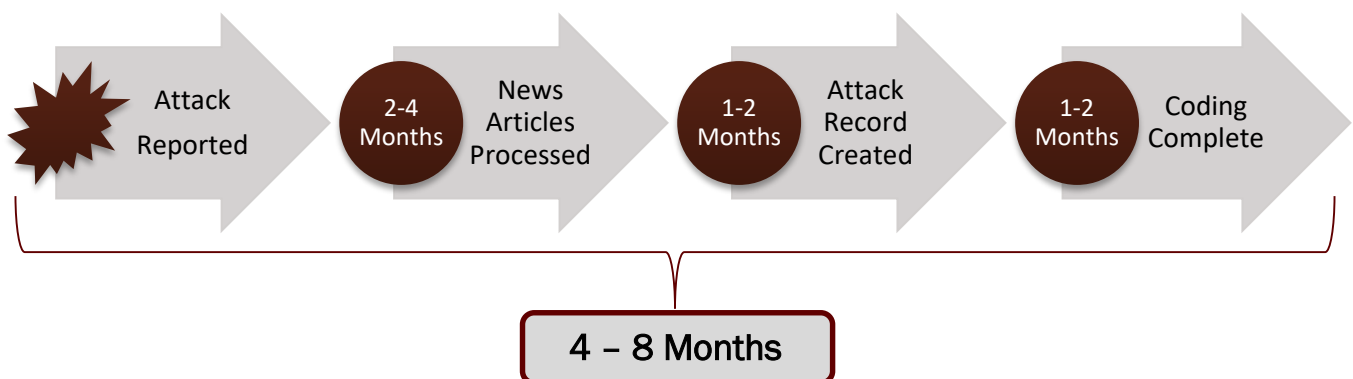


Figure 1 represents the approximate data collection timeline for typical GTD collection. The GTD software architect archives source documents nightly, so when a terrorist attack occurs and initial news reports are published about it, those source documents are maintained for future use. The second arrow represents the pre-processing steps, which usually take place two to four months after the attack. The third arrow represents the triaging process, during which a researcher creates the initial record of the attack, typically one to two months after pre-processing. The detailed coding process to systematically record structured information for more than 100 variables and conduct a final review is completed approximately one to two

months after triaging, represented by the fourth arrow. The premise of the real-time pilot project is that the GTD team could theoretically produce a preliminary version of the dataset based on records produced during the triaging process, prior to the full coding phase of data collection.

Although it captures a broad range, this excludes the best-case and worst-case scenarios for the timeline. For example, with funding from the German Federal Foreign Office and the United States Department of Defense, in 2019 the GTD team reduced the lag time by streamlining the pre-processing steps and increasing triaging capacity. By recruiting and training part-time research assistants to contribute to triaging, there were fewer gaps in triaging during periods when researchers were focused on other tasks including recruiting, training, and mentoring new students. Over the course of the year, this strategy gradually reduced the lag behind real time, until the team was triaging news articles within several weeks of their publication date in late 2019. However, in 2020 the GTD experienced a lapse in project funding that caused data collection to slow significantly and even stop entirely for several months. As a result, the lag time increased rapidly and by March 2021, the research team was triaging source articles that were published in August 2020. Given the shift in focus to the pilot project in 2021, as well as significant turnover in research personnel, by March 2022 the team was triaging news articles published in January 2021.

The research team selected April 2021 to conduct the real-time pilot exercise because 1) it allowed sufficient time to complete the planning stages for the pilot and make necessary updates to the data collection tools, 2) it occurred late enough in the academic semester that students working on GTD collection for 2020 were fully trained and researchers could focus attention on real-time triaging, and 3) it allowed enough time after the real-time triaging exercise to complete the full data collection process for attacks that occurred during the pilot month. On April 5, 2021 the team temporarily stopped triaging news articles published in 2020 to focus efforts on triaging content from April 2021 in real time.

Triaging Collaboration

Triaging potentially relevant source articles published in a particular month is a collaborative task. At the time of the pilot project in April 2021, eight researchers were responsible for triaging. Individual researchers log in to the DMS, review the source documents presented to them, and create initial records of terrorist attacks in the GTD. However, the thousands of news articles published about terrorist attacks around the world are not neatly organized. Multiple researchers encounter a variety of news articles about the same, similar, or related events, and the task of disentangling them requires coordination. The DMS assigns triagers to “clusters” of topically similar articles to review, fencing researchers off from each other so that multiple triagers are not working in the same topical space at once. Along with other measures, this helps prevent duplicate records from being created during the triaging process. It also helps prevent triagers from working independently to clarify a complicated attack, only to find once they establish what happened a colleague who has been working to sort through articles about the same event created the attack record in the interim. However, the clusters are not fool-proof. The GTD team uses a group messaging platform for routine communication and established a “trialoging discussion” channel reserved for messages about triaging. It is not unusual for a triager to post a message to the group such as, “Please steer clear of Afghanistan from October 17, I’m working on three big attacks and a few smaller attacks from the 17th with articles extending into the 18th. The big attacks are in Ghazni, Paktika, and Farah, so you might be alright if you have attacks in other provinces on those days.”

The thousands of news articles published about terrorist attacks around the world are not neatly organized.

Likewise, as the team works through the pool of relevant news articles that need to be triaged, articles that are not needed are discarded. If a news article describes an attack that fails to meet the inclusion criteria for the GTD, or if it offers no new or unique information about a terrorist attack, the researcher removes it from the pool of remaining articles and the rest of the team will not encounter it. There are also situations where some of the media coverage describing the characteristics of an attack suggest it *should* be included in the GTD, but other articles contain information about the same attack confirming it *should not* be included in the database. In these situations, if a triager ultimately decides to exclude the event based on the full set of reports and discards the source articles, a colleague could encounter a lingering article referencing the same attack and decide to create the attack record in the GTD because they did not have access to the sources presenting information that negates that decision. To avoid this, triagers aim to ensure that all articles about the attack that has been excluded are processed, but they might also post a message to the group such as, “I did not create this Indonesia event. FPI supporters opened fire on police officers that were pursuing them. A few articles refer to it as an attack, so watch out for any stragglers.”

The data collection timeline during the pilot project reduced the gap between the attack and completed data from months to days.

Triaging Variables

A key strength of the GTD is the inclusion of more than 100 variables that capture the characteristics of each terrorist attack. However, the initial attack record created during the triaging process normally includes very few. The goal of triaging is to identify attacks for inclusion in the GTD; record the date, location, and a brief summary of the attack in order to disambiguate it from other events; and attach the supporting source documents to the record so that others can fill in details during subsequent stages of the data collection process. The small number of variables allows the research team to complete the triaging process quicker, leaving more nuanced decisions about attack details to the domain-specific coding teams.

Since one of the goals of the pilot project is to inform a strategy for publication of preliminary data, it is noteworthy that triaged records are barely useful for analysis. The event records produced during the triaging phase of data collection could provide information about the number of terrorist attacks by country over time, but they would not support quantitative analysis or modeling of casualties, tactics, perpetrators, or targets. For the purpose of the pilot project, the subject matter team leads identified a small number of key variables from each coding domain to add to the triaging process. The goal of adding these variables, as described below, is to enhance the usefulness of the preliminary data produced during triaging, without slowing down the workflow.

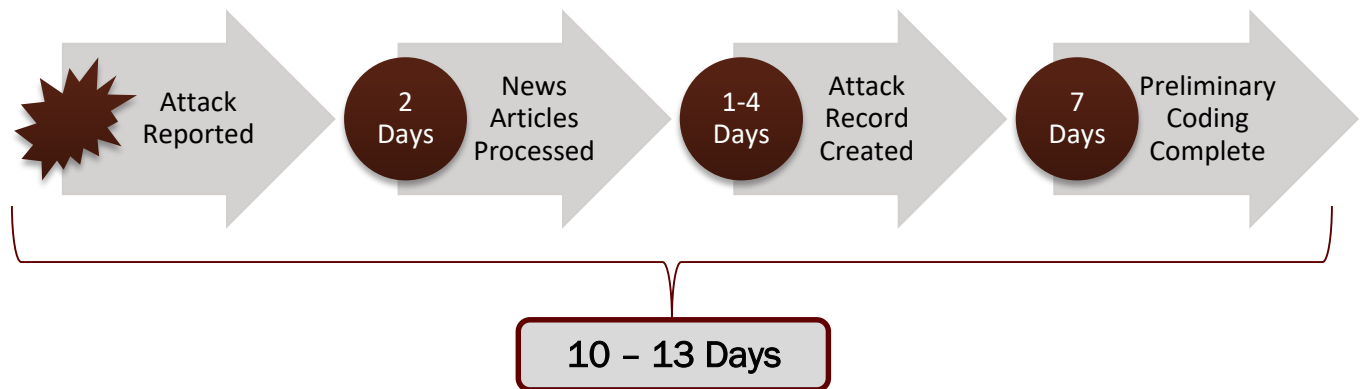
Real-Time Data Collection Pilot: Project Design

Comparison Datasets

To evaluate the reliability of early reporting as it relates to both inclusion/exclusion decisions and the accuracy of attack details, the team collected one month of data—April 2021—in two ways and then compared them. The first method involved adding simplified versions of key variables—those that are normally recorded in more detail later in the workflow—to the triaging phase of data collection. Using a simplified set of key variables allows for important information to be included, without overwhelming and delaying the triaging process by requiring researchers to make numerous difficult coding decisions for each event.

As shown in Figure 2, the data collection timeline during the pilot project reduced the gap between the attack and completed data from months to days. Because the triaging phase of data collection can occur shortly after news articles are published, the research team coded these variables using only sources published within a few days of the attack. Researchers were able to continually edit and update the event record throughout the real-time triaging process.

Figure 2. Approximate Timeline: Real-time Pilot GTD Collection



The second method of data collection uses the conventional GTD workflow described above. After the research team completed the real-time triaging exercise in April 2021, the team triaged source articles published in May 2021 (not in real time), capturing updated information published in May about events that took place in April. Specialized domain coding teams record the details of an event using all available sources, including initial reports and any reports published in the weeks after it took place. With the benefits of hindsight and additional reporting, the data collection team may be able to make more informed decisions

Comparison Datasets

Preliminary GTD Triaging Data - Limited to information collected about April 2021 events based only on news articles published in April 2021 and triaged between April 5, 2021 and May 5, 2021, with nightly snapshots of the database to track when changes were made.

Full GTD Data - Information about April 2021 events documented during the real-time triaging process in April, plus any additional news articles published in May 2021. The records were fully coded by the research team, including domain-specific coding teams and final reviews.

regarding the details of each event. Furthermore, the data files undergo several additional rounds of review as they make their way through the assembly line workflow, presenting opportunities to identify and correct errors. Note that the coding teams were not able to see the preliminary variables documented during the real-time triaging process, as their task was to make independent decisions based on all available information and not be influenced by the preliminary coding decisions.

Ultimately, the pilot produced two versions of a single month of data, allowing for comparative analysis of changes in the attack records made as new information became available and the record was subjected to additional review.

Research Questions

The analysis is guided by a series of questions to assess the timing and substance of the real-time data collection and subsequent updates.

1. How many case records were created?
 - a. What changes were made to event records?
 - i. How many event records were added after real-time triaging was completed?
 - ii. How many event records were deleted during and after real-time triaging?
 - iii. How many event records were updated during and after real-time triaging?
 - b. How many days passed between the initial creation of event records and final updates?
 - c. Are there any noteworthy characteristics of the changes made or of the event records that were changed?
2. What is the accuracy of information recorded during the triaging phase of data collection, compared to the final version of the data? For each information domain:
 - a. Location
 - b. Perpetrators
 - c. Targets
 - d. Weapons/Tactics
 - e. Casualties/Consequences
3. What additional information is informative?
4. What are the implications of real-time data collection for process and efficiency?

Adapting the Triaging Workflow

Based on the subject matter expertise of the specialized domain coding teams and discussion among the triagers, the following variables were included in the pilot for real-time triaging:

Location: A single field for the first-order administrative division where the attack occurred (known as *provstate* in the GTD).

For comparison, the standard triaging interface includes the country where the attack took place and a text field for unstructured location details. The full GTD coding process for locations information includes the country, province/state, city/village, latitude, longitude, specificity of coordinates, and additional location details

Perpetrators: A single field for the name of the perpetrator group(s) believed to be responsible for the attack (known as *gname* in the GTD).

For comparison, the standard triaging interface does not include any structured information about the perpetrators (individuals or groups). It would normally only be referenced in the event summary. The full GTD coding process for perpetrator information includes the names of up to three perpetrator groups believed to be responsible for the attack, the name(s) of individual assailants responsible for the attack, the number of assailants, whether or not the perpetrators claimed responsibility for the attack, the medium/mode for any claims of responsibility, and additional perpetrator details.

Targets: A single, simplified field for the categorical type of target, including Business, Government, Security Forces, Private Citizens/Property, Infrastructure, and Multiple (the full version of this variable is known as *targtype* in the GTD).

For comparison, the standard triaging interface does not include any structured information about the targets or victims of the attack, which would normally only be referenced in the event summary. The full GTD coding process for target information includes the names, organizations, types, subtypes, and nationalities of up to three entities targeted in the attack. The full target type classification scheme includes more than 20 types and more than 100 subtypes.

Weapons/Tactics: A single, simplified field for the categorical type of weapon, including Explosives, Firearms, Other, Multiple and Unknown (the full version of this variable is known as *weaptype* in the GTD).

For comparison, the standard triaging interface does not include any structured information about the weapons or tactics used in the attack. It would normally only be referenced in the event summary. The full GTD coding process for weapons and tactics information includes up to three tactic types (categorical), and the types and subtypes for up to three weapons used in the attack, as well as additional details. The full set of variables in the GTD for weapons/tactics includes eight tactic types, and more than 10 types and 30 subtypes of weapons, plus additional details about the weapons and tactics used.

Casualties/Consequences: One numeric field estimating the total number of people killed, one numeric field estimating the total number of people injured, and one Yes/No field indicating whether hostages were involved in the attack.

For comparison, the standard triaging interface does not include any structured information about the casualties or outcomes of the attack, which would normally only be referenced in the event summary. The full GTD coding process includes extensive information about the consequences of the attack, including the number of people killed, injured, or taken hostage in the attack, as well as the outcome of hostage events, and information about property damage caused by the attack.

Additional Variables: In addition to the domain-specific variable described above, triaging for the project pilot included an additional Yes/No variable for triagers to note an exceptional degree of uncertainty or ambiguity about whether or not the definitional inclusion criteria are met (to prompt further review during analysis), a field capturing the status of the record with respect to source validity requirements, and a field for the analyst to record additional notes during the triaging process.

Adding these variables to the triaging phase of data collection required adapting the triaging interface in the Data Management System (DMS). Figure 3 shows the triaging interface for creating new attack records in the DMS, with the original fields on the left, and the new pilot fields in the gray box on the right. The development of these adaptations was initially completed on March 31, 2021, and the real-time data collection pilot triaging began on April 5, 2021 and ended on May 5, 2021.

Figure 3. GTD Data Management System interface for creating new attack records

GLOBAL TERRORISM DATABASE

Create New Attack

Title	Source	Date	Location	Validity
Kashmore Police foil bid to kidnap two laborers	The Regional Times of Sindh	2020-08-27	undetected	3 secondary

Incident Date

August 2020

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Date Uncertain

Country:

City (Area/Location):

Criteria: 1 2 3

Details/Notes:

Real-Time Pilot Fields

Perpetrator Group:

Weapon:

Target:

Provstate/Admin1:

Min Estimate nKill:

Min Estimate nWound:

Hostages (y/n)

Purgatory:

Extra-uncertainty

Extended Triaging Notes:

Incidents

Afghanistan: Nagorkhel, Nangarhar
08-20
2020: Assaultants attacked D security posts in Nagorkhel, N. At least 4-6 border soldier were killed and two police officer assaultants were injured. Taliban claimed responsibility.

Afghanistan: Lala Gozar din district, Takhar
08-20
2020: Assaultants attacked se Takhar, Afghanistan. At least militia members and two a-6 militia members and three d in the attack. No group clarity for the incident; however, the attack to the Taliban. (No attacks on 3 outposts. Waiting clarify if 2 other attacks show)

Afghanistan: Ghor-Kabul, Firoz Koh, Ghor
08200002

Submit

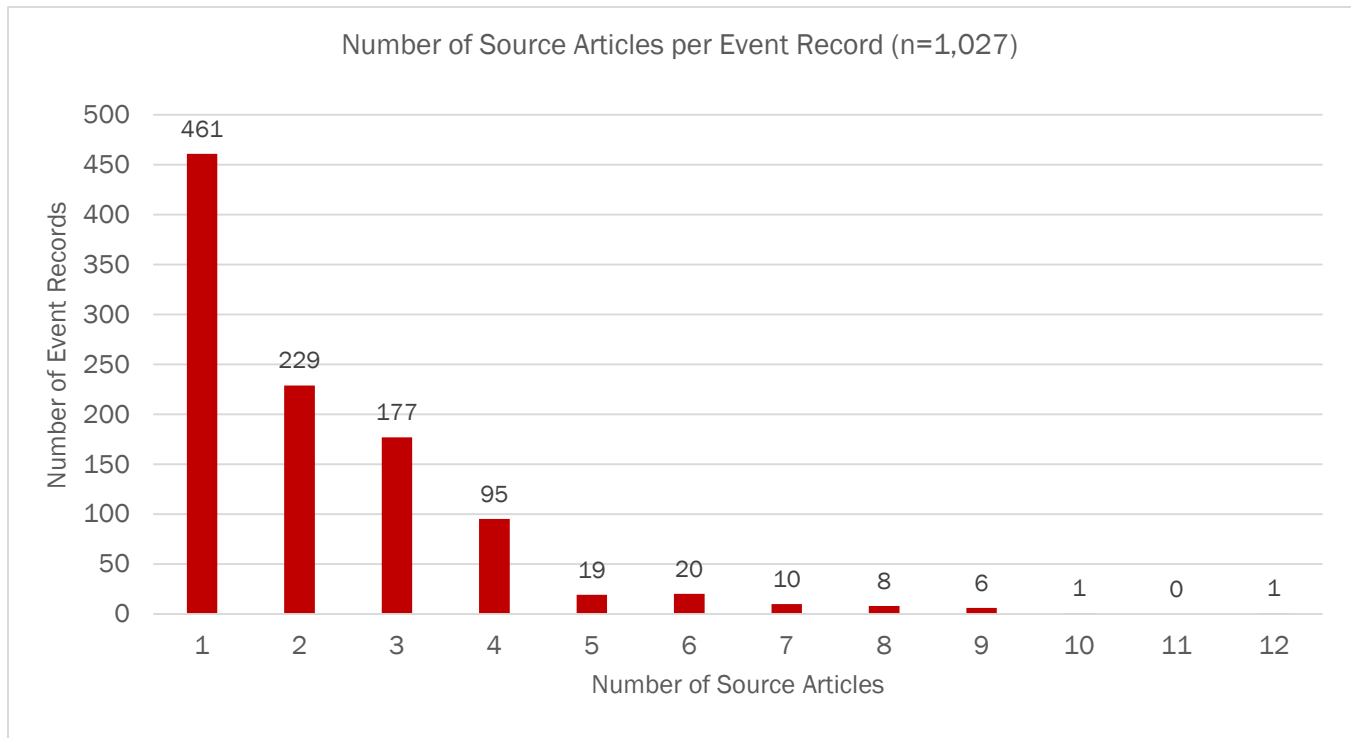
Results

Event Records Created, Added, and Deleted

During the real-time triaging of articles published in April 2021, the research team reviewed 13,747 news articles and created records for 1,084 terrorist attacks. Throughout the data collection process, 57 attack records (5%) were removed for various reasons, leaving 1,027 attack records remaining. This analysis includes any edits made to event records through December 31, 2021.

The research team attached 2,255 news articles to the 1,027 event records as supporting documentation. Shown in Figure 4, the number of articles per event record ranged from one to twelve. Of the 1,027 attack records, 461 (45%) were sourced by only one article. This is consistent with general patterns—45 percent of all GTD event records for 2020 had only one source document attached. Note that this does not necessarily mean that the attack was only reported in a single source. There may have been additional sources published that were duplicative, or otherwise did not provide additional information that was relevant to GTD collection.

Figure 4. Number of Source Articles per Event Record (n=1,027)



There were 31 event records (3%) for attacks that took place in April 2021 but were documented after the real-time pilot triaging ended. These attacks happened in 20 different countries, with no particular indication of geographic concentration. Fifteen of these 31 attacks were non-lethal and six involved an unknown number of people killed. In the deadliest of these attacks, six people were killed. Seven of the 31 attacks involved people taken hostage or kidnapped—it is not uncommon for reports of kidnappings to emerge only after the victim is released or killed.

Two-thirds of the 31 attacks that were documented after the real-time pilot was completed happened in the last week of April, and the source dates for the new attacks are clustered in the first week of May, suggesting that a significant portion of the late additions were a result of the real-time pilot’s arbitrary end date, rather than new information coming to light about earlier events.

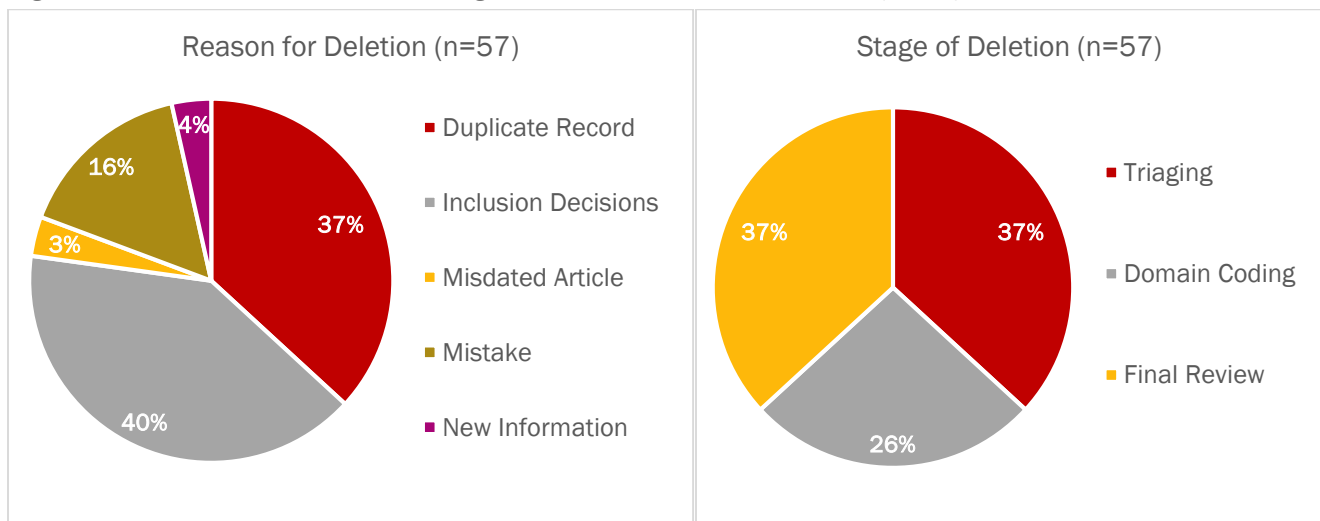
Of the 1,084 event records originally created during the real-time triaging exercise, 57 (5%) were deleted for a variety of reasons, shown in Figure 5. The most common reason for an event record to be deleted is “inclusion decisions,” meaning that upon further review and possibly discussion among the team, it was determined that the attack did not satisfy the inclusion criteria for the GTD. Note that this is distinct from the category of “New Information,” which applies to only two of the deleted records (4%). The remaining reasons for deletions were various types of errors—either a duplicate record had been created, the source article had been misdated, or the triager made a mistake while creating the event record and marked the case for removal.

More than one-third (37%) of the decisions to remove cases were made during the triaging process itself, meaning that if these records were published as preliminary data the errors could likely be identified and

addressed prior to publication. However, 63% of the deleted cases (36 attack records) were identified at later stages of data collection, either during the domain-specific coding process or the final review.

Forty-three event records were marked for “extra uncertainty” during the triaging process—a designation intended to signal a need for additional consideration because the triager believed the early details published about an attack were too ambiguous to confidently make a decision about inclusion. Fourteen of the event records marked as “extra uncertain” were among the records eventually deleted from the database

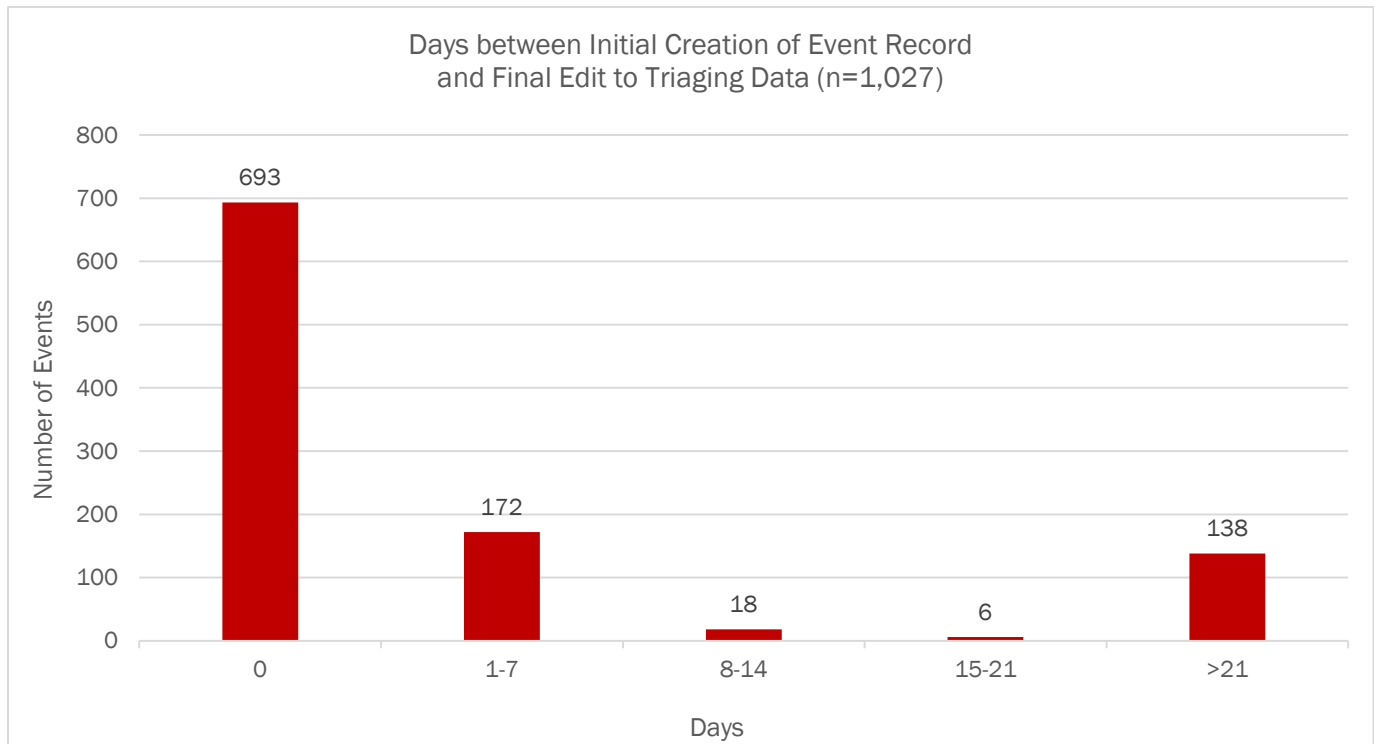
Figure 5. Reason for Deletion and Stage of Deletion for Event Records (n=57)



Event Records Changed

During the real-time triaging of April 2021, the GTD software architect took nightly “snapshots” of the state of the data to track changes made to the data over time. Shown in Figure 6, analysis of these snapshots indicate that two-thirds of the event records created during the pilot period (693 attacks; 67%) were unchanged after the day they were created. An additional 172 events were edited between one and seven days after the event record was created, meaning 84 percent of the event records were essentially stable within a week of being entered into the database. Recall that the triaging process was taking place 3-6 days after the news articles were published so beyond two weeks after an attack, changes to the record produced during triaging were rare. Thirteen percent of event records (138 attacks) were edited more than three weeks after being created, including any changes made after the team triaged the news articles published in May 2021.

Figure 6. Days between Initial Creation of Event Record and Final Edit to Triaging Data (n=1,027)



Understanding the substance of these changes is important. Among the 334 event records that were edited after the day they were created, updates include correcting spelling errors, adding supporting articles, or making more substantive changes regarding the characteristics of the attack. Table 1 shows the types of changes made to event records, the number of records impacted by each type, and that number as a percentage of all records that were changed, and as a percentage of all records that were created (after deletions).

The most common type of change is editing the event summary, as 76% of the event records that were changed involved updated event summaries. This is expected because nearly every other type of change listed would likely warrant an update to the event summary. Two-thirds (68%) of the event records that were changed involved new source documents added, which provides some insight regarding the likelihood that the update was substantive and a result of new information rather than a clarification or technical correction to existing information.

The remaining types of changes were made to a small number of event records—around one-quarter, or less, of the records that were changed and less than 10 percent of the overall number of records created during the pilot. Changes to location information typically involved adding further detail that might help the locations coding team identify the geo-coordinates of the attack. Changes to the province or first-order administrative region were typically spelling standardizations, and in three cases the country where the attack took place was edited. Two were on the border between Saudi Arabia and Yemen, and one country was changed to Nigeria, having been mistakenly left blank initially.

Table 1. Types of Changes Made to Event Records Following Initial Documentation of Attack (n=334)

Type of Change	Number of records changed	% of records changed (n=334)	% of all records (n=1,027)
Event Summary	255	76%	25%
New Source Added	228	68%	22%
Extended Notes	88	26%	9%
Location (Any)	84	25%	8%
<i>Location Details</i>	82	25%	8%
<i>Province/State</i>	9	3%	1%
<i>Country</i>	3	1%	0%
Incident date	51	15%	5%
Perpetrator Group	37	11%	4%
Number of People Killed	31	9%	3%
Number of People Injured	30	9%	3%
Target Type	11	3%	1%
Weapon Type	8	2%	1%
Hostages Y/N	7	2%	1%

Most of the records for which perpetrator group information was updated (29 attacks; 78%) went from unknown perpetrator to known perpetrator information, perhaps as claims of responsibility were published or investigations identified assailants. Eight records that had updated perpetrator group information involved either changing the name of the group responsible to a different group name, adding a second perpetrator group, or changing the spelling or naming convention for the group name.

One-third of the changes to casualty information (21 attacks; 34%) occurred when the number of people killed or injured was initially documented as unknown but later updated with specific information. The most dramatic changes in the number of casualties occurred when a mass casualty attack was determined to be two coordinated events and a single event record was divided into two records.

Preliminary Data vs. Full Data

At the end of the triaging process the research team removed any cases that had insufficient sourcing— inclusion in the GTD requires support from at least one independent, high-validity source. For April 2021, this reduced 1,027 attack records created to 794 attack records to be included in the final data.

By comparing the final snapshot of the preliminary data (taken on May 6, 2021) to the full version of the data that was coded by the domain-specific teams and subjected to additional review, we can assess how “correct” the preliminary data is. Because the variables documented during the real-time triaging exercise were simplified compared to those in the full GTD codebook, it was necessary to establish some rules to define “correctness.” For example, the six categorical options for target type presented to triagers can be aligned with the 22 categorical options for target type in the GTD codebook. In most cases these alignments were straightforward, but some situations required additional review. In general, the intent for this measure of correctness was to focus on substantive accuracy. Preliminary values that were substantively accurate were classified as correct even if there were, for example, difference in spelling for a province or perpetrator group name.

Table 2 shows the number and percentage of event records that were correctly coded in the preliminary data collected during the real-time triaging exercise. The variable that performed best was the province or first-order administrative region where the attack occurred, with 97% accuracy, followed by the question of whether hostages were taken, which matched the full dataset 96% of the time. The perpetrator group information was accurate in 95% of all cases, including 35 records (4%) where changes were made to conform to naming conventions for consistency. While not impactful regarding substantive accuracy, these types of data cleaning issues do impact the usability of the data. It is possible that improved auto-completion tools can help reduce these types of discrepancies, however that would not account for situations like the emergence of new groups and it may introduce opportunities for error if coders make assumptions about the name presented by auto-completion being the correct name, even if the name presented in sources is slightly different.

Table 2. Number and Percent of Event Records Coded Correctly in the Preliminary Data (n=794)

Variable	# Coded Correctly During Triaging (n=794)	% Coded Correctly During Triaging (n=794)
Province/State	770	97%
Hostages Y/N	762	96%
Perpetrator Group	757	95%
Number of People Killed	710	91%
Number of People Injured	707	89%
Weapon Type	670	84%
Target Type	603	76%

Information about casualties recorded during real-time triaging was accurate for 91% of attack records with respect to deaths, and 89% of attack records with respect to injuries. There is a larger drop-off in accuracy for weapon type (84%) and target type (76%). This is not especially surprising, given that these are variables that have numerous categories in the GTD codebook. However, there are some technical issues related to the framing of these variables that could partly explain the poor accuracy and offer opportunities for improvement. For example, 124 of 191 “incorrect” targets classifications (65%) were for cases that were ultimately classified as having multiple targets in the full dataset, but the triager did not select the “multiple targets” entry. Likewise, 49 of 124 “incorrect” weapons classifications (40%) were for cases that were ultimately classified as having multiple weapons, but the triager did not select the “multiple weapons” entry.

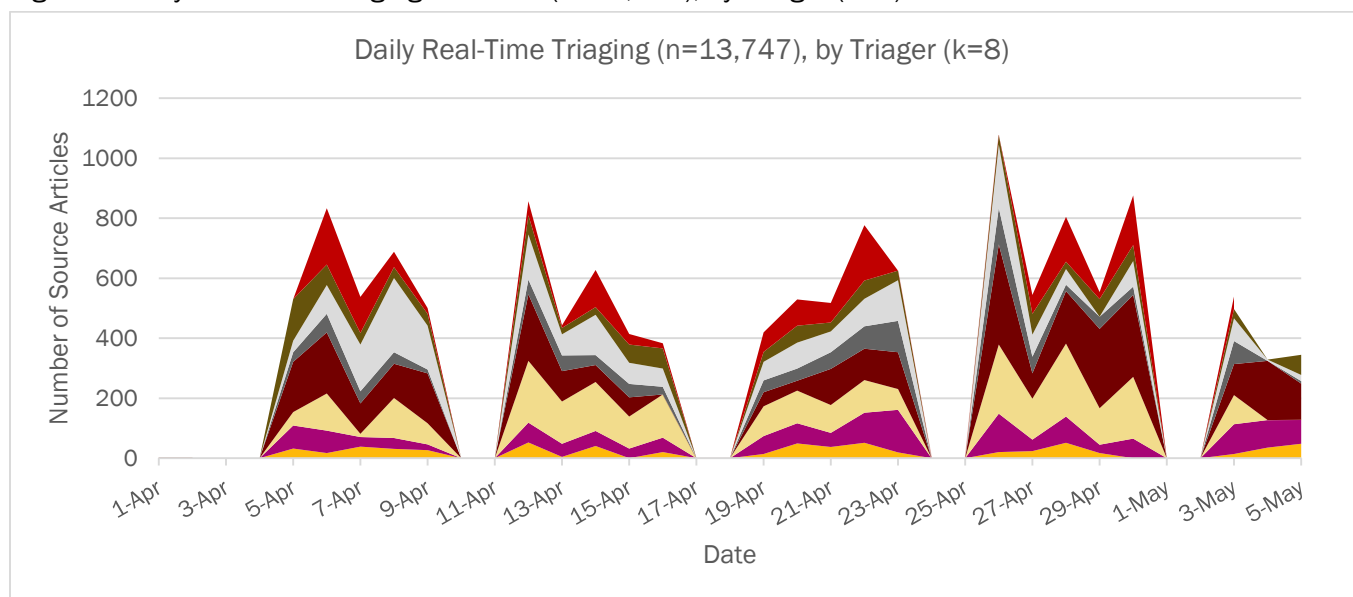
This suggests that these preliminary variables are particularly challenging for complex attacks that involve multiple targets or weapons. This may be because additional information about weapons or targets was published later. However, given how rarely updates to these variables appeared in Table 1, it is more likely a result of triagers not fully capturing the nuances of weapon use and targets associated with the attack at the time they created the event record. There may be better ways to structure this information in the preliminary data, but it is useful to consider these variables more vulnerable to inaccuracy or inconsistency when using preliminary data for analysis.

Process/Efficiency

Although the main goal of the pilot project was to compare the information collected in preliminary real-time data to fully vetted information, the experience of designing and executing the pilot project yielded important lessons about challenges related to the process and efficiency of real-time data collection for the GTD. First,

and most importantly, it is not sustainable under the circumstances of the pilot project. During the real-time triaging pilot, the eight researchers responsible for triaging focused nearly 100% of their effort on triaging. Likewise, the GTD software architect performed pre-processing and maintenance tasks nightly and each weekend. Figure 7 shows the number of source articles triaged daily by the team, broken down by triager, with each colored band representing a different individual researcher. Normally, these researchers would spend 50–75% of their time on triaging, but divide their remaining time between various tasks, including training and mentoring students, reviewing completed data for consistency and accuracy, and specialized research on topics related to terrorism to help inform decision making. The task of triaging more than 13,000 news articles about violence and keeping pace with incoming source documents in real time is mentally taxing and could not be done by a team of eight researchers on an ongoing basis.

Figure 7. Daily Real-Time Triaging Statistics (n=13,747), by Triager (k=8)



Second, triaging in real time requires extraordinary collaboration and communication among the team. As noted above, the GTD team uses a group messaging platform with a dedicated channel for “trialog discussion” to deconflict efforts and make decisions about how to handle challenging cases. Although the triaging discussion channel had been in operation since 2016, when the team concluded real-time triaging for April 2021, 25% of the overall messages exchanged had been posted that month. To minimize overlapping efforts, triagers are assigned to clusters of news articles when they log in to the DMS. The researchers on the GTD team work very well together, and the generally positive assessment of the preliminary data presented above is undoubtedly a reflection of the team’s focus, skill, and dedication to documenting events as accurately as possible. However, the topical space is quite narrow when working on news articles published in a period of a few days, compared to a full month. This requires a significant effort to avoid triaging traffic jams. Somewhat paradoxically, increasing the number of triagers to reduce the workload for individual researchers would likely exacerbate issues related to coordination and avoiding “stepping on others’ toes.”

Finally, real-time triaging is cognitively inefficient. When triaging news articles several months after they were published, the reader can review the news coverage holistically, looking at documents published the day of the attack and in the days that followed to get a sense for what happened and how the information

developed over time. After synthesizing this information and selecting the most useful source documents to support the event record, the triager can remove all unneeded source documents and move on. In comparison, when triaging in real time a researcher encounters articles about a given attack as they are published. They review the articles published the day of the attack and make a decision, only to start over the next day and the day after that, each time refreshing their recollection of decisions made previously and adjusting as needed. This can be challenging even when the researcher who encounters the information about the attack on day two is the same person who handled it initially. But it is more likely that a different researcher encounters the information the next day, needing to recreate the situational awareness that supported their predecessor's course of action. In addition to being repetitive, this often means that the event record unnecessarily includes supporting articles that are not the most informative overall, but happened to be the most informative articles available on the day they were triaged. This creates downstream implications, where the research team either needs to actively remove obsolete, low-quality source articles. Or, anyone coding or reviewing the event record later must sift through them to find what they are looking for.


Conclusions

In general, the real-time pilot project was successful, providing valuable information about the stability and accuracy of early reporting and the logistics of real-time data collection. The analysis of event records created and updated during the real-time triaging process indicates that when data collection takes place within 10 to 13 days following a terrorist attack, the substantive details are highly stable. Many attack records (67%) remained unchanged after the day they were initially documented, and most (84%) remained unchanged as of one week after the record was created—approximately two weeks following the attack. The updates made to the remaining 16 percent of events were often technical corrections rather than substantive developments.

It is important to acknowledge that the lack of updates in most cases does not necessarily mean that the information documented for those events is perfectly accurate—it is certainly possible that new or corrected information about a relatively low-profile attack did not result in updated media publications. The conclusion from these statistics is simply that the attack record is likely as complete as possible, given the research team's access to information.

Depending on the particular use of the data, changes to 16 percent of cases could have a significant impact. Certain variables, including the date and location of the attack, perpetrators, and whether the attack involved hostages, were quite reliable. Information on casualties was also correct in approximately 90 percent of attacks. Other variables, such as classification of weapons used and targets attacked, were much less comprehensive before they were reviewed by subject-matter experts focused on those domains. The results suggest that in most cases these updates to event records were not a result of new information being published, but instead reflected the difficulty of consistently and thoroughly capturing weapon and target information for complex attacks.

The results are generally encouraging for the potential usefulness of preliminary GTD data. When “real time” is measured in days, rather than hours, the research team can identify and document valuable information about terrorist attacks during the triaging process. The remaining stages of GTD data collection play an important role in promoting completeness, capturing complexity, and standardizing information for ease of use, but for certain modeling tasks the preliminary data would likely be sufficient. However, real-time triaging also has practical implications with respect to sustainability and efficiency that would need to be addressed



in order to be feasible. Several of these challenges—like those related to managing workloads, situational awareness, and cognitive burdens—would be improved upon by performing data collection within a few months of attacks happening, rather than a few weeks. A longer time lag would allow a larger group of analysts to work concurrently without overlapping efforts in a small topical space. It also means the reporting for nearly all attacks will have tapered off, so researchers can quickly focus on the most informative, accurate sources and avoid repetitive analysis of poor source articles.