Cyber Events Database Codebook

Charlie Harry, Nancy Gallagher, Bill Lucyshyn and Devin Entrikin

March 2026 (Revised to incorporate GDELT NGRAMS and GDELT Article List as primary data sources)

Note: This change involves adding GDELT's vast, real-time monitoring of global news articles (covering over 60 languages and millions of sources) to identify candidate cyber events through keyword searches. GDELT focuses on open web news sources, but the core purpose, fields, and validation processes remain essentially unchanged.

Thurgood Marshall Hall,
School of Public Policy, University of Maryland
College Park, MD 20742
(301) 405-7601

SCHOOL OF
PUBLIC POLICY
CENTER FOR GOVERNANCE OF
TECHNOLOGY AND SYSTEMS

Cyber Events Database

PI: Charles Harry, PhD
Co-PI: Nancy Gallagher, PhD
Program Manager: Devin Entrikin

Purpose

The increasing scale and impacts of cyber events remain an enduring concern. Yet, information covering the range of threat actors, motive, industry, or classified impact is scarce, fractured, or is only available through private organizations at a significant cost. The Cyber Events Database collects publicly available information on cyber events from 2014. It was created to address a lack of consistent, well-structured data necessary for making strategic decisions about how to invest resources to prevent and respond to cyber events. The Cyber Events Database allows users to distill analytical insights on cyber threats to specific industries and regions, trends over time, and the behavior of different threat actors. By leveraging GDELT's Web News NGrams 3.0 and Article List datasets, the database draws from a comprehensive, real-time global news monitoring system to enhance coverage and timeliness of cyber event data.

Data Collection Method

Data is collected using a mixed-methods approach that integrates a Python script (using internal libraries csv, datetime, urllib, requests, json, and external library Beautiful Soup 4) to scrape data from known open internet and dark web sources, accessing main landing pages and RSS feeds via predetermined URLs to extract date published, title, URL, article preview, local date/time of access, and overarching website title in HTML format. Concurrently, the supplemental script leverages the GDELT Project's Web News NGrams 3.0 Dataset, performing keyword searches (e.g., "cyber attack," "data breach," "ransomware") on real-time unigrams and higher n-grams from millions of news articles across over 60 languages, joining matches with the GDELT Article List to retrieve article URLs, publication dates, titles, previews, and metadata (e.g., language, source domain). All data is processed into two .csv files—one for scraped web data and one for GDELT-derived data—which are deduplicated daily and reviewed by researchers to (1) ensure events meet the definition of a cyber event (a single or cumulative unauthorized effort using computer technology and networks to achieve a discernible effect on a target), (2) categorize threat actor type, motive, threat actor country, and targeted country, and (3) classify affected industry and effects based on a structured taxonomy[1]. Attributions are taken directly from source material without independent validation.

The script, fully functional in Python 3.8 or higher, supports querying GDELT's historical data from 2014 onward with real-time updates every 15–60 minutes, though GDELT integration into this CEDB begins in CY 2025.

[1] Harry, C., & Gallagher, N. (2018). Classifying cyber events. *Journal of Information Warfare*, 17(3), 17-31.

The script makes no effort to determine the suitability of the candidate for the cyber event. All linked entries/articles are included in a daily deduplicated file to be reviewed by a researcher, who makes final judgments as to whether events are valid members of the dataset. We define a cyber event as the result of any single unauthorized effort, or the culmination of many such technical actions, that engineers, through the use of computer technology and networks, achieve a desired primary effect on a target. The dataset chiefly records individual cyber events where a discernible effect was achieved by the threat actor (e.g., hacker). To be included in the dataset, each event must be traced back to an underlying source describing details surrounding the event itself.

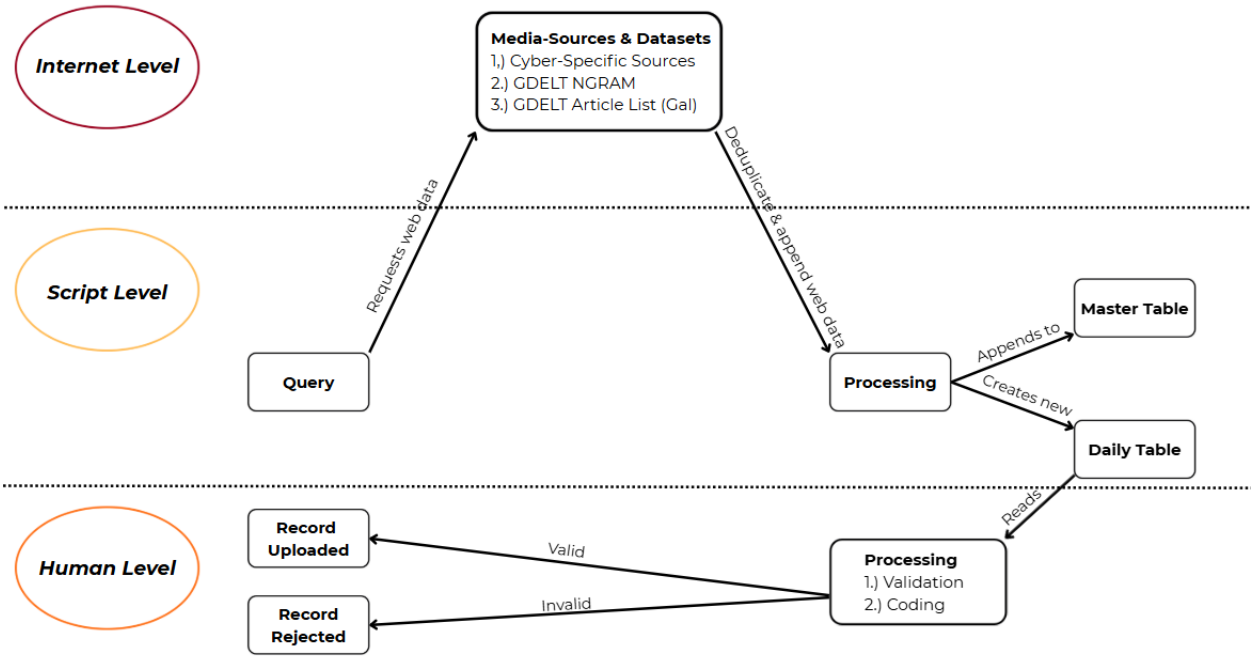

Figure 1. Data Collection Process

Fields and Description

- Unique identifier (*slug)* – a unique 16-digit alphanumeric identifier assigned to each record

- Identify events without GDELT *(original_method*) – To allow researchers to control for differences in data collection methods when analyzing trends across the entire dataset (2014–present) and account for the shift to GDELT in 2025 an Integer (0 or 1) will indicates whether the event data was collected using the original web scraping method (pre-GDELT, before January 2025) or the GDELT-based method (January 2025 onward).

- Event Date (*event_date*) – Date or estimated date that the event occurred in DD-MM-YYYY format. Estimated dates are accurate to the month and are indicated as the first day of that month. If no date is identified, this field will match the *reported_date* value.

- Date of publication (*reported_date)*: The date, in DD-MM-YYY format, the media source published the article identifying the relevant cyber event.

- Year (*year*) – Year event occurred in YYYY format.

- Month (*month)* – Month event occurred in MM format.

- Actor (*actor*) – String variable indicating the name of the organization or individual responsible for the event; “undetermined” if unknown.

- Actor Type (*actor_type*) – Categorical variable indicating the nature of the actor responsible for the event:
    - Criminal – Organization that illicitly accesses networks for financial gain.
    - Nation-State – A government agency, military, or affiliate thereof
    - Terrorist – a non-state actor seeking to influence political or military conditions by targeting civilians
    - Hacktivist – an individual or group motivated by social or political activism
    - Hobbyist – an individual motivated by curiosity or prestige

- Organization (*organization*) – String variable indicating the name of the target organization whose networks were illicitly breached

- North American Industry Classification System (NAICS) Code (*industry_code*) – Two-digit NAICS code defining the target organization.

- Industry Name (*industry*) – String variable indicating the name of the NAICS code category

- Motive (*motive*) – Categorical variable indicating the intended results sought by the actor committing the event
    - Protest – The disruption of services in order to send a political or social message to the target organization, or to a government or population indirectly.
    - Sabotage – The intentional, irreparable destruction of information, networks, or devices
    - Espionage – Accessing of networks for the purposes of intelligence or surveillance.
    - Financial – Exfiltrating sensitive data for direct or indirect financial gain.

- Event Type (*event_type*) – Categorical variable indicating whether the primary end effects of the event were disruptive, exploitative, or a mixture of the two.
    - Disruptive – Impedes the target organization’s normal operations
    - Exploitive – Illicitly access or exfiltrate sensitive information such as personal identifiable information, classified information, or financial data.
    - Mixed – Event incorporates both disruptive and exploitative elements, such as a ransomeware attack.

- Event Sub-type (*event_subtype*) – Categorical variable further classifying the nature of an event based on the part of the target organization's IT infrastructure that was most seriously impacted, regardless of the tactics or techniques used to achieve the final result.

    - Disruptive events:
        - Message Manipulation – Interference with the target organization's ability to accurately present or communicate information to its customer base, constituency, or other audience.
            - Examples: These attacks include the hijacking of social media accounts, such as Facebook or Twitter, or defacing a company website by replacing the legitimate site with pages supporting a political cause.

        - External Denial of Services – Executed from devices outside of the target organization's network to degrade or deny its ability to communicate with other systems.
            - Examples: Many types of Distributed Denial of Service (DDoS) attacks would fit into this category, including ICMP flood, SYN flood, or ping of death. A Border Gateway Protocol (BGP) hijack that diverted Internet traffic away from a targeted organization's website would also fit in this category.

        - Internal Denial of Services – Executed from inside a target organization's network to degrade or deny access to other parts of the IT network.
            - Examples: An attacker who gained remote access could move laterally inside an organization's network to reset a core router to factory settings, preventing devices inside the network from communicating with each other. They could also install malware on a file server and disrupt data sent to and received from user workstations.

        - Data Attack – The manipulation, destruction, or encryption of data in a target organization's network.
            - Examples: Common techniques include the use of wiper viruses and ransomware. Using stolen administrative credentials to manipulate data and violate its integrity, such as changing grades in a university registrar's database, would fall into this category, as well.

        - Physical Attack – The use of IT components, such as SCADA systems, to manipulate, degrade, or destroy physical systems.
            - Examples: Current techniques used to achieve this type of effect include the manipulation of Programmable Logic Controllers (PLC) to open or close electrical breakers, leading to a de-energizing of that portion of the grid, or the utilization of user passwords to change settings in a human machine interface so that a blast furnace overheats and is destroyed.

- Exploitive events – exploitative events are classified by the part of the target organization's IT infrastructure from which the malicious actor steals the information.
    - Exploitation of Sensors – The theft of data from a peripheral device, such as a credit card reader, smart TV, or baby monitor.
        - Example: In 2013, the Target corporation had thousands of their Point of Sale (PoS) devices compromised, leading to the loss of over 40 million customer credit card numbers.

    - Exploitation of End Host – The theft of data stored on user's desktop computers, laptops, or mobile devices.
        - Examples: Common tactics currently used include sending a malicious link for a user to click or leveraging compromised user credentials to log in to an account.

    - Exploitation of Network Infrastructure – The theft of data through direct access to networking equipment such as routers, switches, and modems.
        - Example – In 2018, over 500,000 routers worldwide were infected with VPNFilter malware which maintained access to devices through the compromise of user credentials and left open the potential for information to be hijacked.

    - Exploitation of Application Server – The use of a misconfiguration or vulnerability to gain access to data in a server-side application (e.g. a database) or on the server itself.
        - Examples: The hacker in the 2015 Office of Personnel Management data breach used a SQL injection to access millions of records with sensitive information about current and former government employees. This category also includes the theft of data from Sony Pictures achieved when the hacker gained direct access to an e-mail server with organizational correspondence.

    - Exploitation of Data in Transit – The acquisition of data moving between devices.
        - Example: Unencrypted data might be acquired as it is sent from a PoS device like a credit card reader to a database, or when somebody makes a purchase over the Internet from their laptop through an unsecured wireless hotspot at a local coffee shop.

- Severity measures for disruptive events (*magnitude, duration, scope)* – Qualitative and quantitative information that describes the magnitude, duration, and/ or scope of the cyber event.

- Severity measures for exploitive events (*ip, org_data, cust_data)* – Qualitative and quantitative information describing the type i.e., intellectual property, organizational, and/or customer, and amount of data compromised.

- Event Description (*description*) – String variable consisting of 1-3 sentences detailing the event.

- Source URL (*source_url*) – String variable consisting of the URL from which the data was pulled. String variable consisting of the URL from which the data was pulled, typically sourced from the GDELT Article List.

- Target Country (*country & country_iso3*) – String variable consisting of country name and 3-letter ISO country code for the target organization's location.

- Actor Country (*actor_country & actor_country_iso3*) – String variable consisting of country name and 3-letter ISO country code for the actor's location.

- U.S. State and County (*state; county)*: String variables that identify the sub-national geographic location of a cyber event when the target organization is located in the United States and sufficient information is available.

- Target Country Organization –A series of binary (dummy) variables in the database schema to indicate membership in the following organizations: NATO, EU, Shanghai Coop, OAS, Mercosur, AU, ECOWAS, ASEAN, OPEC, Gulf Coop, G7, G20, AUKUS, CSTO, OECD, OSCE, and Five Eyes. Each field will indicate whether the target country is a member of the respective organization (1 = member, 0 = non-member).

Contact Information

Center for Governance of Technology and Systems (GoTech)
Thurgood Marshall Hall
7805 Regents Drive
College Park, MD 20742
Phone: 301-405-7601

Charles Harry, PhD: charry@umd.edu
Nancy Gallagher, PhD: ngallag@umd.edu
Devin Entrikin: dentik@umd.edu