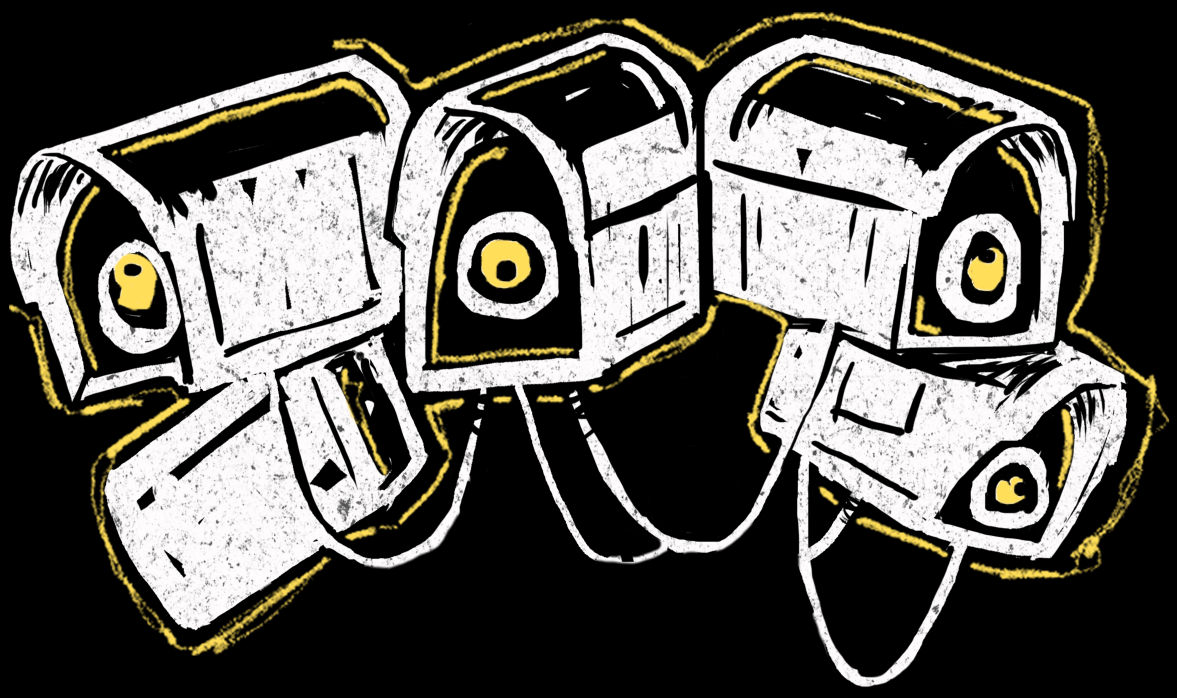
TRUSTLAB



Digital Defence Playbook

Table of Content

DEFENDING DEMOCRACY

A playbook for detecting and exposing digital subversion campaigns, specifically built for human rights defenders tracking conversations that would subvert democracy or otherwise harm citizens. This is a nonpartisan resource for detecting and exposing influence operations by domestic and foreign state actors, as well as non-state actors, ranging from political activists to paid lobbyists, conspiracists, and extremist agitators.

Fighting propaganda with propaganda normalises information warfare, further polarises our societies, and erodes public trust in all information. This playbook is, therefore, explicitly intended for civic self-defence and does not offer or advocate offensive counter-measures.

Credits:

The playbook is distilled from a decade of accumulated knowledge and tradecraft pioneered by Code for Africa's (CfA) information integrity teams, working across 28 African countries.

- Chief strategists: **Justin Arenstein**
- Editors: **Amanda Strydom (South Africa)**
- Lead researchers: **CC. Chargi, Nirali Patel (Kenya) and Dr. Sandra Roberts (South Africa),**
- Contributing editors: **Jacktone Momanyi (Kenya)**
- Copyeditors: **Gloria Aradi, Mary Mutisya and Mwende Mukwanyaga (all from Kenya)**
- Designers: **Temidayo Oyegoke (Nigeria)**
- Project coordinators: **Jones Baraza (Kenya) and Wairimu Maina (Kenya)**

Editions:

This English edition of the playbook was produced by CfA through the TrustLab consortium, comprising CfA, Deutsche Welle Akademie, and Siasa Place, with funding from the European Union.

GLOSSARY

This is a comprehensive list of terms and abbreviations used throughout the document, organised in alphabetical order. The lexicon is internationally accepted by the European External Action Service, the International Fact-Checking Network, and the United Nations (UN).

AI	Artificial intelligence (AI) is a set of technologies that enable computers to perform a range of functions that often simulate human intelligence.
CBO	Community-based organisation
CSO	Civil society organisation
Cyberbullying	The deliberate sending, posting, or sharing of negative, harmful, false, or mean content about or to a victim.
Digital threat	A digital threat is any malicious or manipulative activity conducted through digital platforms or systems to disrupt, deceive, or harm individuals, institutions, or societies. An example is the use of malware, such as viruses.
DISARM	The Disinformation Strategic Analysis, Response, and Mitigation (DISARM) framework is a model for analysing the tactics, techniques, and procedures (TTPs) used in disinformation campaigns. It involves four phases: planning, preparation, execution, and evaluation, with countermeasures at each stage.
Disinformation	False or inaccurate information intentionally spread to mislead and manipulate people, often to make money, cause trouble, or gain influence.
Hate speech	Communication that attacks or uses pejorative or discriminatory language against an individual or group based on their identity, ethnicity, or religious affiliation.
Information manipulation	This is the deliberate distortion, omission, or fabrication of information to influence public perception, behaviour, or decision-making.
Misinformation	False, incomplete, inaccurate, or misleading information or content generally shared by people who are unaware that it is incorrect or deceptive.
NGO	Non-governmental organisation
OSINT	Open source intelligence (OSINT) is the collection, processing, and analysis of publicly available information.
SOCMINT	Social media intelligence (SOCMINT) is a sub-branch of OSINT that involves the collection, analysis, and interpretation of data from social media platforms.
Surveillance	The monitoring of individuals or groups, often by government agencies, to gather information, influence behaviour, or maintain control.

How to use this playbook

This section provides guidance on how to navigate and use this playbook.

Getting started

Before starting any threat detection activities, ensure your team understands basic digital security. Use secure communication channels and protect sensitive data, especially when monitoring political content or hate speech.

Equipment requirements

You need a computer or smartphone with reliable internet access. A laptop or desktop computer works best for data analysis and dashboard creation, but basic threat detection can begin with a smartphone. Ensure you have sufficient data allowance or Wi-Fi access for regular online research and platform monitoring.

Software requirements

The playbook does not need you to subscribe to costly services. It uses Google Workspace apps for most of the examples, but you can use alternatives if you have them. You will need Google Sheets or an equivalent spreadsheet for data organisation and analysis.

A web browser supports most monitoring activities through platforms such as MediaCloud and social media sites. Google Drive provides secure evidence archiving and team collaboration. The playbook also references Google Alerts for keyword monitoring and Google Forms for stakeholder data collection. It may recommend additional tools or apps that offer free versions.

Expertise levels

This playbook works best with teams that combine different skills. Different sections require different expertise, so you can assign team members to techniques that match their strengths.

- **Basic research skills:** Stakeholder mapping and basic threat landscape analysis work well for team members who are comfortable with spreadsheets, internet searches, and social media navigation. These sections do not need specialised technical knowledge.

- **Data and technical skills:** MediaCloud monitoring and lexicon building suit team members familiar with search operators such as (AND, OR, NOT) and basic data analysis. Assign these to people who understand CSV files and can create simple visualisations.

- **Investigation skills:** Kill chain analysis and network mapping work best with team members who understand investigation frameworks. Some sections reference tools such as Gephi for network analysis, though step-by-step guidance is provided. Experience in research methodology and report writing is beneficial.

- **Language skills:** Assign monitoring tasks to team members who are fluent in relevant local languages. Understanding cultural nuances, political terminology, and how harmful language manifests in different contexts directly impacts success. Local political knowledge is crucial here.

Select the sections you need

Each section operates independently. While threat landscape analysis is recommended as the starting point for frontline human rights defenders setting up early warning systems to detect and expose illicit influence campaigns, you can focus on specific sections based on your needs. If you already have monitoring systems and solely need hate speech lexicons, skip directly to that section. If you need stakeholder mapping for a specific project, use that section.

How each section is structured

Each section follows the same format – introduction and purpose, key concepts, step-by-step processes, roles and responsibilities, required tools, best practices, and success metrics. This consistent structure helps you quickly find the information you need.

- **Introduction and purpose:** This section provides a brief overview of the activity. It explains what this section aims to achieve (objectives), outlines its scope, and sets expectations for how it will support consistent execution and continuous improvement.

- **Concepts and terminologies:** These are important to users as they help them understand the service and how it works. Additionally, they provide clarity on the language being used throughout the document.

- **Processes and workflows:** This section defines clear and precise steps for achieving the task at hand. It includes a step-by-step approach, highlighting every task that needs to be undertaken.

- **Roles and responsibilities:** This section presents suggestions on the 'who' and 'what they do' based on the activity provided. Identifying roles reduces the opportunity for overlaps and gaps, ensuring accountability.

- **Tools and resources:** This section lists any relevant tools and resources required to perform the activities provided.

- **Best practices:** These are proven methods and techniques implemented to help organisations leverage successful strategies and mitigate challenges.

- **Metrics and performance indicators:** These help track progress and outcomes, which are essential for continuous improvement.

TRUSTLAB

Chapter 1
Introduction

1.Introduction

Covert manipulators increasingly subvert democratic processes using sophisticated digital techniques to deceive, polarise, and incite the public for political gain. Illicit influence operations have become a common tactic in hybrid warfare, undermining trust in government institutions and destabilising countries.

State-affiliated strategic communications (StratCom) professionals, ideological campaigners, and freelance 'keyboard warriors' drive these operations, working for a range of clients. Digital mercenaries rapidly adapt their tools and techniques to evade civic defenders.
This playbook equips human rights defenders with early warning systems to detect and expose illicit influence campaigns. It introduces key aspects of information disorder and resilience, highlighting strategies to safeguard information integrity, support free speech, and combat information pollution.

Users will learn how to map information ecosystems, monitor key actors, identify threats, and track content to protect the integrity of information flows. The playbook outlines a methodology, along with tools and resources, to enable CSOs or development agencies to replicate research and analysis processes most applicable to their circumstances.
The core focus is to build an early warning system to disrupt threats.

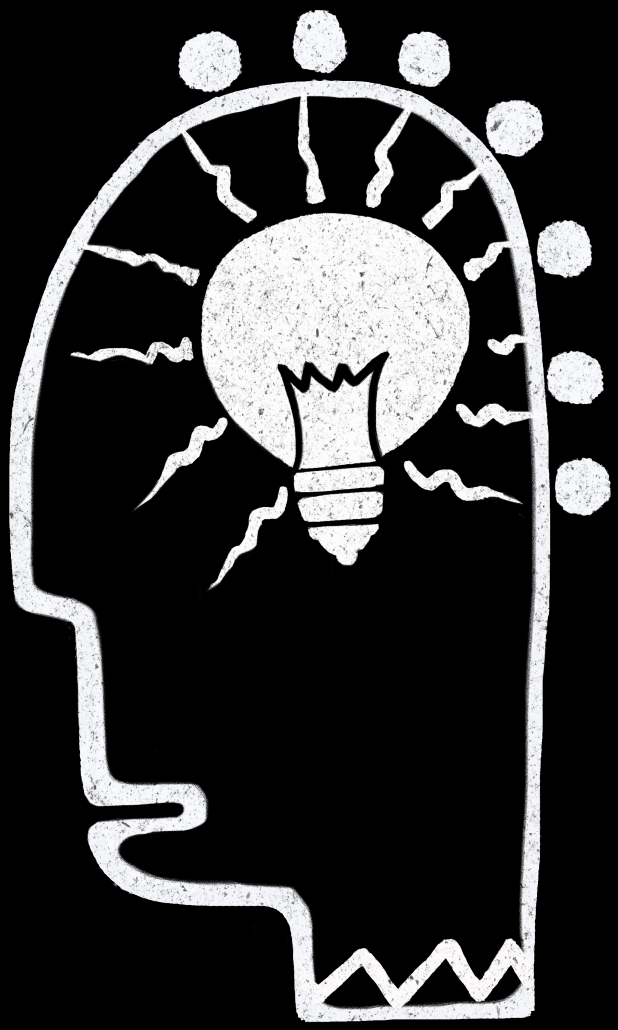
The document further outlines methods to conduct threat landscape analysis, map stakeholders, develop cyber kill chain strategies, and build lexicons and watchlists. It also covers how to set up real-time dashboards to track digital news media on TrustLab's MediaCloud toolkit, as well as tips for producing threat reports and other evidence dossiers that regulators or technology platforms can act on.

Practical guidance on team structure, workflow, and security ensures that those working in information resilience operate efficiently and safely. Secure communication, information management, and the amplification of insights are essential to drive meaningful action in the fight for information integrity.

A living online toolkit complements this playbook, offering updated resources and tools. A network of rapid-response investigators from the African Digital Democracy Observatory (ADDO) and the African Fact-Checking Alliance (AFCA) supports these efforts, alongside operational security assistance from the TrustLab.

This playbook serves as a foundational resource. Civic defenders seeking to enhance their skills can access modular, step-by-step training through the ADDO Academy. For a more advanced version, please visit the [add link to the sheffield playbook}

TRUSTLAB



Chapter 2
Threat landscape analysis

2.Threat landscape analysis

2.1 Introduction

Threat modelling is a systematic approach to identifying, assessing, and mitigating digital threats that may affect a target organisation or community. In the context of this playbook, threat modelling will focus on digital harms, including surveillance, cyberbullying, disinformation, hate speech and malicious AI

.

The objective of this section is to provide a structured and practical approach to mapping digital threats that influence ordinary citizens, CBOs, and CSOs. By the end of this process, the analysis will identify specific digital threats, the actors involved, and the most vulnerable communities.

2.2 Concepts and terminology

- **Amplification:** This refers to the process by which content is spread or 'made viral'.
- **Dark web:** Refers to the hidden parts of the internet that are accessible only through special tools and often used for anonymous communication or illicit activity.
- **Dark social:** Refers to untraceable content sharing via private channels such as messaging apps or email, making it difficult to monitor the spread of disinformation.
- **Fact-checking infrastructure:** These are organisations and tools used to verify and debunk false claims.

2.3 Processes and workflows

A baseline understanding of an existing digital environment is required to start an analysis.

- **Step 1: Country context analysis**
  Establish a foundational understanding of the target country's information ecosystem. This step includes examining the media landscape, levels of internet and social media penetration, popular search trends, and the robustness of the local fact-checking infrastructure. Public reports, government datasets, and insights from social analytics platforms support this analysis.
- **Step 2: Open web assessment**
  Assess the broader digital environment, focusing on the legal frameworks governing information flows, the influence of advertising-driven content, and the visibility or marginalisation of certain narratives in mainstream media. This process involves reviewing policy documents, civil society publications, and media monitoring.
- **Step 3: Open social media analysis**
  Analyse activities on open social media platforms, looking at both regulatory approaches and findings from fact-checkers to uncover how mis-/disinformation circulates. This approach can blend qualitative insights, such as narrative framing, with quantitative data such as volume, virality, and engagement metrics.

- **Step 4: Dark web and dark social exploration**
  Identify content circulating on less accessible platforms such as Telegram, WhatsApp, and private forums, which are often used to evade public scrutiny. Apply OSINT techniques such as keyword tracking, manual group discovery, cross-platform link tracing, and the use of bot analysis tools to monitor public and semi-public spaces on these platforms. This procedure helps identify recurring risks, thematic patterns, and the potential for narratives to migrate across platforms.

- **Step 5: Threat actor TTPs**
  Document observed TTPs used by actors spreading disinformation. These TTPs are categorised under tactics such as evasion (e.g., inauthentic accounts), manipulation (e.g., deepfakes or sentiment seeding), and amplification (e.g., coordinated posting or bot networks). Real-world examples should support each tactic where possible.

Table 1; below lists examples of TTPs used in the four phases of an influence operation.

Phases	TTP	Description
Plan	Determine target audiences	Identify the audiences most vulnerable to manipulation.
	Leverage existing narratives	Build upon culturally relevant or pre-existing beliefs and stories.
Prepare	Create inauthentic social media pages and groups	Set up fake accounts, pages, or groups to build an initial audience or community.
	Use copypasta	Copypasta is a block of text repeatedly copied and shared online, and sometimes modified over time as users edit, add, or remove parts before reposting the text.
Execute	Flooding the information space	Overwhelm the information environment with volume to dilute truthful content.
	Share memes	Spread digestible, emotionally charged content for rapid virality.
Access	Content focused	Analyse how specific narratives or formats perform online.
	Behaviour changes	Evaluate if the content resulted in attitude or behaviour shifts.

2.4 Roles and responsibilities

a. These are the skills required to conduct a threat landscape analysis:
b. Make sense of local stories and explain the context behind them in a way that supports deeper analysis.
c. Find and verify open-source info (OSINT) from platforms like social media, news sites, forums, and search engines.
d. Track engagement and trends by processing basic metrics, following search trends, and creating simple visuals to support your findings.
e. Organise and structure reports so they're clear, consistent, and match the organisation's style and goals.
f. Know the laws around your research topics, especially how legal frameworks define things like hate speech.

2.5 Tool and resources

Table 2; below presents the tools essential for performing a threat landscape analysis and explains the purpose of each:

Tool	Purpose
OSINT platforms a. Google Trends b. Maltego/Vortimo/WHOIS	• Analyse public interest over time based on keyword searches. • Investigate networks, domain patterns, and actor attribution.
SOCMINT platforms a. Meltwater or Meta Content Library (MCL) b. WhatsApp/Telegram monitors	• Track social media and online article trends. • Understand closed-group dark social dynamics.
Data visualisers a. Excel b. Flourish c. Gephi d. Tableau	• Visualise data, build graphs and summarise statistics.

2.6 Best practices

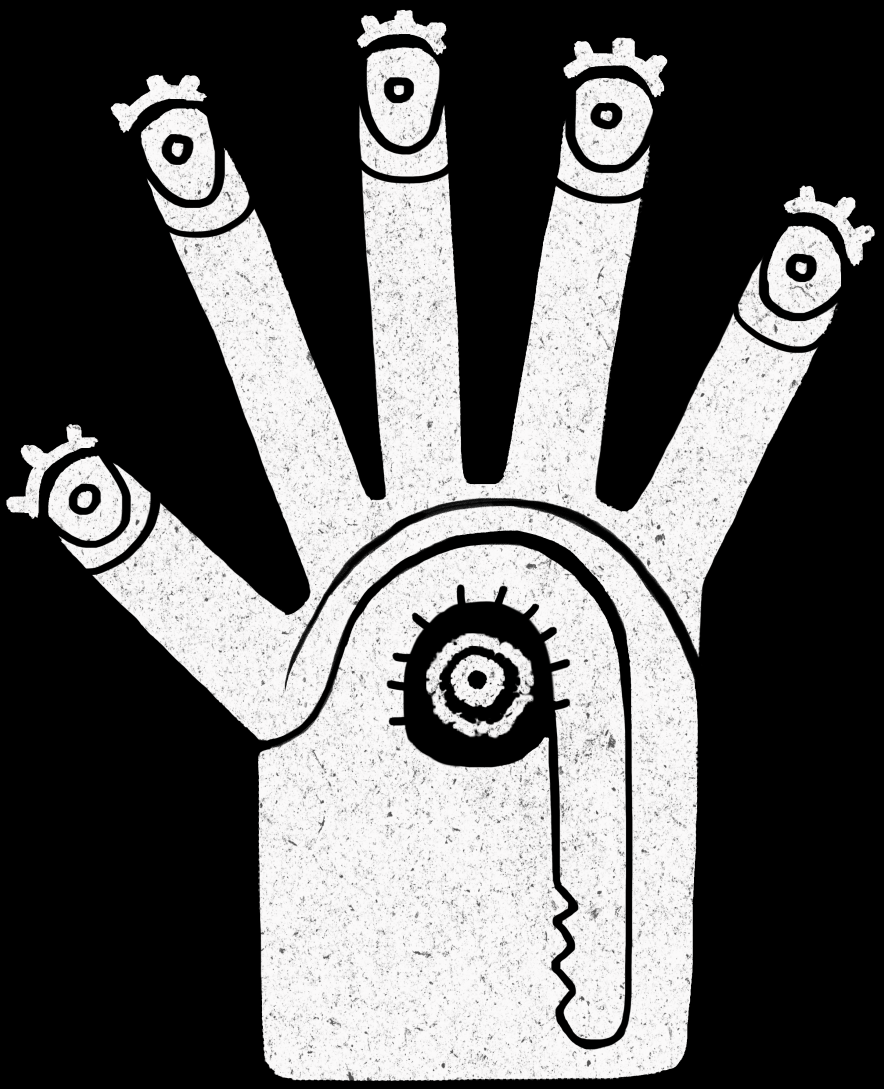
i. Cross-check everything: Use social media, mainstream media, and fact-checking sites to verify narratives and spot patterns.
ii. Tap into local knowledge: Work with civil society groups, journalists, and researchers to decode cultural context and nuance.
iii. Connect the dots: Track how online content spills into real life, from an X post to a protest or a policy shift, to understand the actors, impact, and risks.
iv. Keep a paper trail: Log your search terms, sources, and research steps to make your work transparent and reproducible.
v. Save your receipts: Use digital archiving tools to preserve the evidence you gather.

2.7 Metrics and performance indicators

Table ;The following are important indicators to keep track of during research and report writing for the threat landscape analysis.tt

Metric	How to measure/track
Reach and engagement of manipulated information spread by bad actors	• Volume of mentions and spread across platforms, number of views, shares, and likes from social monitoring tools.
Report uptake	• Includes stakeholder feedback, downloads, or policy citations.
Accuracy of insights	• Cross-validation with fact-checkers and real-world developments.
Team workflow efficiency	• Timeliness of delivery, collaboration, and feedback.
Reader understanding	• Internal or external surveys and informal feedback sessions.

TRUSTLAB



Chapter 3
Stakeholder mapping

3.Stakeholder mapping

3.1 Introduction and purpose

Stakeholder mapping is a systematic process that identifies and analyses individuals, groups, and organisations with significant interest or influence. Stakeholder mapping is crucial for deciding communication priorities and managing risks to projects.

This section aims to provide a structured approach to identifying relevant actors, prioritising engagement efforts, and establishing collaborative relationships.

3.2 Concepts and terminologies

- **Stakeholder:** An individual, group, or organisation that can affect or be affected by your project. This includes those who can support or influence your project outcomes.
- **Stakeholder groups:** Categories of stakeholders may include beneficiary communities, CSOs, funding partners, local businesses, local government entities, media outlets, national government departments, and traditional leaders.
- **Influence level:** The amount of power or impact a stakeholder has in relation to your project. It is categorised as high, medium, or low.
- **Interest level:** The degree to which a stakeholder cares about your project goals. It is categorised as high, medium, or low.
- **Snowball sampling:** A technique where initial stakeholder contacts help identify additional relevant stakeholders.

3.3 Processes and workflows

This section provides a step-by-step approach to stakeholder mapping for a country or region of interest. Each main step includes specific subtasks that should be completed in sequence to ensure thorough and effective stakeholder identification, engagement, and documentation.

Step 1: Identify potential stakeholders

a. Conduct social media research

i. Search Facebook using specific terms related to your area of interest (e.g., 'women group Mombasa').
ii. Search X (formerly Twitter) using regional hashtags combined with thematic tags (e.g., #WomenEmpowerment #HomaBay).
iii. Join Facebook groups where community organisations share updates.
iv. Participate in relevant WhatsApp groups where organisations of interest network.

b. Perform internet research

i. Use specific search terms that combine locations and focus areas (e.g., 'community-based organisation Lamu' + 'digital security').
ii. Check local government websites for directories of registered community groups.
iii. Search for county-specific NGO directories.
iv. Look up 'grassroots organisations' in combination with county names and thematic areas.

c. Access the umbrella organisation information

i. Review online membership directories of networks and coalitions in your sector.
ii. Access publicly available member lists on umbrella organisation websites.
iii. Check annual reports of umbrella organisations for partner listings.
iv. Review sector-specific coalition websites for participant information.
v. Obtain attendance lists from recent conferences and workshops.
vi. Review speaker and panellist lists from relevant forums.
vii. Check social media event pages for participants.
viii. Review published proceedings from sector events.

d. Access and review official registries and other stakeholder databases

i. Contact departments associated with social security, social development, or gender departments at the county level. The names of departments may differ among counties.
ii. Request access to their databases of registered organisations.
iii. Review the categorisation by location and focus area.
iv. Note the registration status of organisations.
v. Search for reports published by other agencies and development partners.
vi. Access resource directories compiled by research institutions.
vii. Review online platforms that aggregate civil society information.
viii. Check university research repositories for sector analyses.

e. Follow funding trails

i. Identify one active organisation in your target area as a starting point.
ii. Review their 'partners' or 'donors' section on their website.
iii. Navigate to donor websites to find their grantees.
iv. Download and review donor annual reports for comprehensive grantee lists.

Step 2: Make contact and expand your network

a. Select appropriate communication channels

i. Use WhatsApp as the primary contact method for grassroots organisations.
ii. Send emails for more formal organisations and communications.
iii. Consider Facebook Messenger for organisations active on that platform.

b. Implement snowball sampling

i. Ask each organisation to identify others working in similar areas.
ii. Request to be added to relevant WhatsApp groups.
iii. Ask for direct introductions to partners.
iv. Enquire about the formal or informal coalitions they belong to.

c. Collect standardised information

i. Customise the Google Form (https://bit.ly/3FrXtfl) to your needs.
ii. Gather basic details such as organisation name, acronym, and registration status.
iii. Document the primary contact person and their role.
iv. Record the main thematic areas and target beneficiary groups.
v. Map geographic coverage at both county and sub-county levels.
vi. Note all digital presence points, including social media handles.

d. Create and maintain a database

i. Develop a spreadsheet with one row per organisation.
ii. Ensure consistent column headers for all key information.
iii. Include complete contact details and geographic coverage.
iv. Create columns to track engagement status and communication history.
v. Schedule regular updates to maintain accuracy.

e. Classify stakeholders

i. Evaluate interest level (categorised as high, medium, or low) based on mission alignment.
ii. Assess influence level (categorised as high, medium, or low) based on reach, reputation, and resources.
iii. Categorise by organisational type (e.g., CBO, government entity, network, or NGO).
iv. Identify priority stakeholders based on an interest-influence matrix.

Step 3: Create a visual stakeholder map

a. Develop a visual to help conceptualise priorities

i. Place stakeholders in appropriate interest-influence quadrants.
ii. Use visual markers to indicate stakeholder relevance to project objectives.
iii. Determine communication priorities. The focus should be on organisations that are highly interested in and influential to the success of your project.

b. Document engagement strategies

i. Assign specific engagement approaches for each stakeholder quadrant.
ii. Note key actions required for priority stakeholders.
iii. Establish contact frequency based on stakeholder importance.
iv. Assign a team member responsibility for each stakeholder.
v. Set timelines for engagement activities.

c. Share and review the map

i. Present the stakeholder map to your team.
ii. Gather feedback and additional information.
iii. Make revisions based on new insights.
iv. Schedule periodic reviews to update the map as relationships and stakeholders evolve.

3.4 Roles and responsibilities

In a small CBO, stakeholder mapping is a collaborative effort in which all staff members contribute based on their unique knowledge and connections. These roles can be adapted to your organisation's structure, with some individuals potentially covering multiple responsibilities.

a. Director/executive director

i. Provides strategic direction for the mapping process.
ii. Leverages existing high-level connections with other organisations.
iii. Offers insights into the broader NGO landscape based on experience.
iv. Approves priority stakeholders for deeper engagement.
v. Opens doors to networks and coalitions through personal connections.

b. Project coordinator/manager

i. Oversees the day-to-day stakeholder mapping activities.
ii. Ensures alignment with project objectives and timelines.
iii. Coordinates team efforts to prevent duplication.
iv. Maintains the big-picture view of stakeholder relationships.
v. Identifies gaps in stakeholder coverage.

c. Programme staff

i. Contribute sector-specific knowledge about relevant organisations.
ii. Identify stakeholders through their professional networks.
iii. Provide insights on the relevance of various stakeholders to specific projects.
iv. Help assess stakeholder interest and influence levels.
v. Support follow-up engagement with technical stakeholders.

d. Communications officer

i. Serves as the central point of contact for stakeholder outreach.
ii. Manages communication channels such as email, social media, and WhatsApp.
iii. Maintains a stakeholder contact information database.
iv. Crafts appropriate messages for different stakeholder groups.
v. Monitors stakeholder responses and engagement.

e. Administrative support

i. Assists with data entry and database maintenance.
ii. Organises stakeholder information into accessible formats.
iii. Schedules meetings and engagement activities.
iv. Helps prepare materials for stakeholder interactions.
v. Supports documentation of the mapping process.

Remember that in a small CBO, staff often wear multiple hats. The person handling communications might also manage social media research, while programme staff might take responsibility for specific geographic areas or thematic groups. Regular team meetings for sharing information and insights are essential to identifying and appropriately engaging all relevant stakeholders.

3.5 Tools and resources

Table 3; below outlines the relevant tools and resources for stakeholder mapping:

Tool	Purpose
Google Forms to collect stakeholder details	• Collect structured stakeholder information.
Social media platforms	• Research and identify potential stakeholders.
Spreadsheets	• Track stakeholder details and engagement status.
Stakeholder mapping template	• Visualise stakeholder relationships, set engagement priorities for the project, and get organisational buy-in for stakeholder management.
WhatsApp roups	• Network with local organisations.

3.6 Best practices

c. Share and review the map

i. Here are some proven methods and techniques:
ii. Begin research with specific search terms related to both geographic location and thematic areas.
iii. Prioritise WhatsApp and email for communication with grassroots organisations.
iv. Use existing stakeholders to identify new ones through snowball sampling.
v. Maintain structured data collection to ensure consistent information.
vi. Create visual maps that show relationships among stakeholders.
vii. Categorise stakeholders by both interest and influence to prioritise engagement.
viii. Join relevant community groups and networks to build relationships.
ix. Review donor websites and annual reports to identify connected organisations.

3.7 Metrics and performance indicators

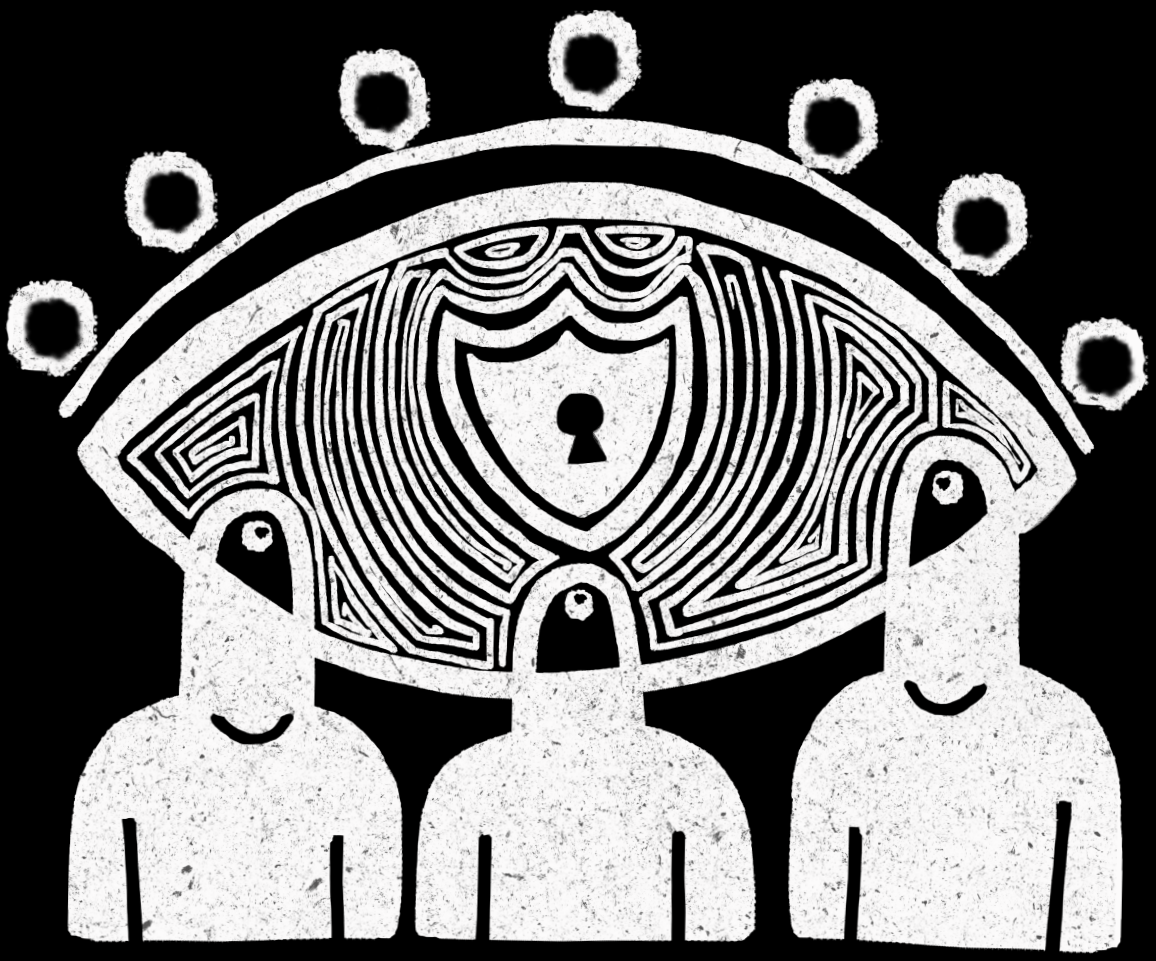
Track the essential metrics outlined in table four below to evaluate the effectiveness of your stakeholder mapping:

Metric	How to measure/track
Stakeholder identification	• The number of stakeholders identified and documented in your database.
Engagement rate	• The proportion of stakeholders successfully contacted, expressed as a percentage of the total number identified.
Representation balance	• A breakdown of stakeholders by major categories, such as CSOs, government, and private sector entities.
Participation rate	• Records of stakeholder attendance and contributions during project meetings and related activities.
Relationship quality	• Assessment of stakeholder satisfaction based on post-engagement feedback collected through standardised 1–5 scale surveys.
Strategic coverage	• Percentage of high-interest, high-influence stakeholders engaged.
Referral effectiveness	• The number of new stakeholders discovered through existing contacts.
Data quality review	• The number of new stakeholders discovered through existing contacts.
Implementation review	• A monthly review of these metrics, accompanied by documented action points to address gaps or areas for improvement.

Implement regular review meetings to assess these metrics against your targets and ultimate project goals and adjust your approach as needed.

Track these metrics through database monitoring, engagement records, feedback surveys, and periodic team reviews of the stakeholder mapping process.

TRUSTLAB



Chapter 4
Developing cyber kill chain strategies

4. Developing cyber kill chain strategies

4.1 Meta's online operations kill chain

4.1.1 Introduction and purpose

The online operations kill chain is a structured framework developed to break down, analyse, and disrupt malicious online campaigns. Whether it is a disinformation network, scam operation, or influence campaign, most of these operations follow a similar set of TTPs. This guide helps teams:

a. Systematically map the lifecycle of an operation.
b. Identify points of detection and disruption.
c. Foster collaboration between investigators across different sectors.
d. Improve defensive strategies over time through repeatable, structured analysis.

4.1.2 Concepts and terminologies

- **Kill chain:** A sequence of steps or tactics a threat actor uses during an operation.
- **Assets:** Controlled resources, such as accounts, emails, domains, malware, etc.
- **Engagement:** Attempts to reach, deceive, or harm a target audience.
- **Longevity:** Actions taken to prolong an operation or avoid takedown.

4.1.3 Ten phases of the kill chain

The kill chain has ten main phases, each representing a major action that threat actors often take during an influence or disinformation campaign. These major actions (called tactics) can be broken down into smaller steps (called techniques), and then into detailed examples (called procedures). For instance, if the tactic employed is 'creating fake accounts', the techniques utilised might include using AI-generated images or copying real profiles.

- **Acquiring assets:** At this stage, malicious actors could create fake email addresses, buy or register new web domains, or set up new social media accounts.

- **Disguising assets:** The actors make the assets look legitimate through actions such as using AI-generated profile pictures, writing professional-looking bios, or impersonating real people or media outlets.

- **Gathering information:** Bad actors research their targets, often using social media scraping, spy tools, bots that monitor hashtags or keywords, or search for local political or social tensions to exploit.

- **Coordinating and planning:** They plan how and when to launch the campaign by using encrypted messaging apps such as Signal or Telegram, or meeting in private Facebook groups/pages or Discord servers.

- **Testing defences:** This step involves trying out different ways to see what actions social media platforms will flag. For example, threat actors might post two versions of the same message (known as A/B testing) to check which one avoids detection. They may also push the boundaries to learn what kind of content is removed or flagged, helping them fine-tune future messages.

- **Evading detection:** Once they know how systems work, threat actors use tricks to stay hidden. This could mean using virtual private networks (VPNs) to hide their locations or writing in coded or vague language that human moderators or AI tools might miss. They might disguise links or use altered spellings and symbols to sneak around filters.

- **Indiscriminate engagement:** Here, the actors try to spread content as widely as possible without worrying about who sees it. It can involve spam posts, mass tagging, or sending the same message to many accounts or groups. The goal is often to create noise or saturate a platform with a message to manipulate trends or visibility.

- **Targeted engagement:** This step is quite focused. Threat actors may send messages to specific people or groups, use targeted ads, or pose as someone trustworthy to trick a target. This approach is often more personal and persuasive and aims for a stronger impact with fewer messages.

- **Compromising assets:** Here, bad actors try to take over real accounts or systems by sending phishing emails or malware or tricking individuals into sharing their login information. Once inside their accounts or systems, they can spread misinformation from trusted accounts, making it seem more believable.

- **Enabling longevity:** Even after being detected and removed, threat actors do not stop; they rebuild. They might create new fake personas, register new websites, or shift to different platforms. This step ensures the operation can keep going despite takedowns or bans, and it often shows planning and resilience behind the campaign.

4.1.4 Processes and workflows

Below is a step-by-step guide on how to use the kill chain:

a. Gather all evidence

i. Collect OSINT or internal data
Start by collecting any material or clues related to the suspected campaign. These include social media posts, screenshots, web links, news articles or blog posts, archived content (using digital archives such as the Wayback Machine), or any data logs or flagged communications. This gives you a foundation to begin mapping out the operation.

Use OSINT tools and sources or any internal monitoring you have access to. Look for:

- **URLs** – Web links being shared repeatedly or suspicious shortened links.
- **Social media accounts** – Unusual or newly created accounts pushing a narrative.
- **Platforms** – Where is the activity happening? (e.g. Facebook, Telegram, WhatsApp, or X)
- Internet protocol (IP) addresses or website domains – If available, these can help identify the origin of a campaign.

Also note patterns, such as:

- Accounts posting the same content at the same time (coordination).
- Content with unnatural language or repetitive phrasing (automation).
- Sudden spikes in engagement (artificial amplification).

ii. Note any suspicious content or behavioural patterns, such as repetitive narratives or hashtags, use of emotional or inflammatory language, fake profiles (e.g., no profile photo, generic bios, or few followers), or engagement that seems 'too fast' or 'too perfect' (indicating possible bot activity).

b. Document TTPs at each phase
Example: The 'disguising assets' phase documents use of Generative Adversarial Network (GAN), generated avatars, and domain mimicry.

c. Compare across operations
Identify if the same TTPs are used in other cases (e.g., GAN use across multiple networks).

4.1.5 Roles and responsibilities

Actions involved in using the Meta online operations kill chain include:

a. Systematically gather raw data from platforms or monitoring tools and prepare it by removing duplicates, errors, or irrelevant entries.
b. Identify irregular spikes in engagement, unusual posting times, or repetitive patterns that may indicate coordinated or suspicious activity.
c. Search for the same posts, links, or user handles across multiple platforms (e.g., Facebook, Telegram, and X) to spot replication or coordinated spreading.
d. Examine account bios, posting patterns, and interactions to assess whether an account is likely inauthentic or part of an influence operation.
e. Identify and label specific behaviours or TTPs (e.g., fake account creation or content flooding) that match known threat tactics.
f. Recognise and log digital traces such as IP addresses, suspicious domains, and malware-related URLs.
g. Connect the technical data to wider patterns, such as reused domains across fake accounts or botnets.

4.1.6 Tools and resources

Table 5; below lists tools and resources required to create and use a kill chain.

Tools	Purpose
OSINT platforms, for example: a. VirusTotal b. Wayback Machine c. Maltego	• Test the visibility of malicious content • Find deleted or hidden evidence • Visualise networks and personas
SOCMINT platforms, such as: a. Meltwater b. Brandwatch	• Monitor narrative spread, sentiment, and amplification across social media platforms

4.1.7 Best practices

a. Not every phase will appear in every case; use only what is observed.
b. Focus on early-stage patterns (asset acquisition, disguise).
c. TTPs evolve; regularly revisit and revise mapping.

4.1.8 Metrics and performance indicators

Table 6; below outlines the performance and effectiveness metrics used to evaluate Meta's online operations kill chain.

Metric	How to measure/track
Number of kill chains completed	• Per operation or campaign
Time to detect (TTD)	• From asset creation to detection
Number of reused TTPs identified	• Across different threat actors
Response time	• Time from detection to takedown/mitigation
Stakeholder feedback	• Surveys, debriefs, and cross-institutional calls

4.2 DISARM framework

4.2.1 Introduction and purpose

This section introduces the DISARM kill chain, which models the lifecycle of disinformation campaigns in four phases: planning, preparation, execution, and evaluation. It aims to provide a structured framework for analysing influence operations, identifying recurring tactics, and informing targeted interventions. By mapping disinformation activities to each phase, analysts can disrupt threat actor workflows and support iterative improvements in detection and response.

4.2.2 Concepts and terminologies

Understanding the following concept is essential:

- **Blue-team tactics:** Defensive or mitigative actions.

4.2.3 Processes and workflows

Below is a step-by-step guide to using the DISARM kill chain:

a. Start by organising what you are seeing into the four big stages of an influence operation. These stages help show where in the process the activity is happening.

i. In the early stage, actors set goals, pick targets, and decide how to influence them. **Example:** Talking in private chat groups or drafting a narrative.
ii. In the second stage, the actors gather tools and build their infrastructure. **Example:** Creating fake social media accounts or registering websites.
iii. The third stage is public-facing, where the actual influence operation is launched. **Example:** Posting misleading content, running a hashtag campaign, or using bots to amplify a message.
iv. After the campaign runs, actors assess how effective it was. **Example:** Measuring engagement or tweaking tactics based on what worked.

b. Identify TTPs: Look for specific behaviours or methods used in the campaign. DISARM provides a taxonomy, a kind of dictionary of known tactics, to help you identify and label them. For example:

i. **T0011:** Compromise legitimate accounts - This involves hacking or taking over legitimate accounts to distribute misinformation or damaging content.
ii. **T0119:** Cross-posting - Posting the same message to multiple internet discussions, social media platforms or accounts, or news groups at the same time to increase the chances of content exposure.

c. Identify how the parts of the operation are connected. Sometimes, one activity enables or supports another.

For example: The creation of fake accounts in the preparation phase allows for amplification of content in the execution phase. Or, the use of closed messaging apps for coordination helps plan how content will be distributed later.

d. After mapping the operation, you can figure out what actions to take and when. Each phase of the kill chain offers an opportunity to disrupt the operation. These are called countermeasures or blue-team tactics.

Examples:
Monitor known threat actors and channels.
Report or block fake accounts or domains.
Use content moderation or fact-checking to reduce the spread.

Analyse data to improve your own detection systems.

4.2.4 Roles and responsibilities

Activities involved in building and applying the DISARM framework include:

a. Map narrative activities to DISARM phases and identify patterns.
b. Identify source content and verify authenticity.

4.2.5 Tools and resources

Table 7; below lists the tools and resources required to build and use the DISARM framework.

Tools	Purpose
DISARM framework	• Phase-based campaign modeling
OSINT platforms, for example Maltego or Vortimo	• Source tracing and actor mapping
SOCMINT platforms, such as Meltwater OR MCL	• Volume and spread analysis

4.2.6 Best practices

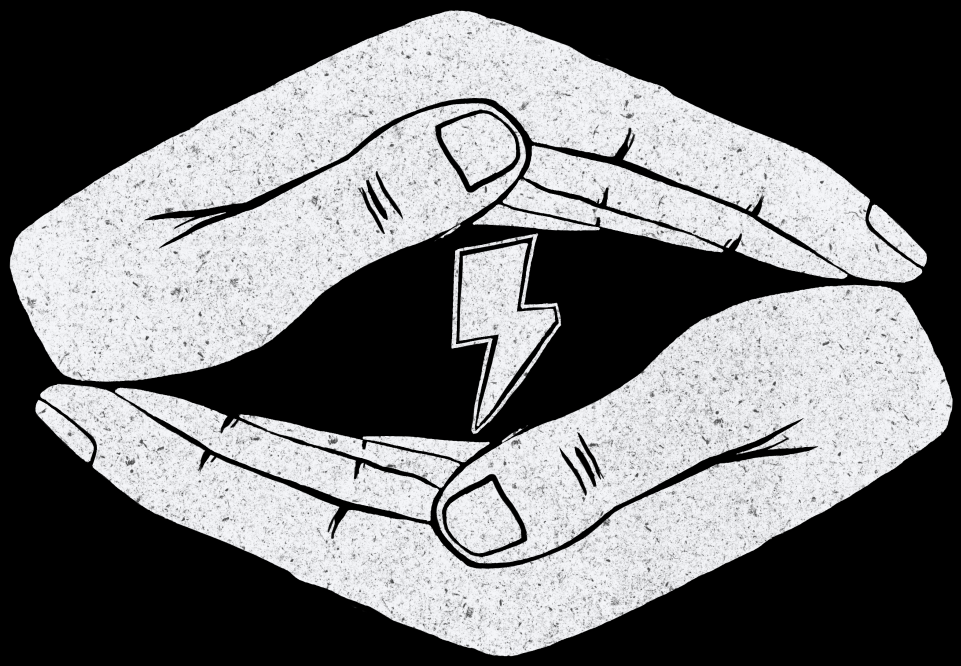
a. Always triangulate between narrative themes, engagement data, and actor behaviour.
b. Use historical cases to anticipate future narrative evolutions.
c. Avoid over-attributing without verified links to actors or platforms.

4.2.5 Metrics and performance indicators

Table 8; below outlines the main metrics used to evaluate the effectiveness of applying the DISARM framework. The metrics help in gauging how well disinformation threats are being detected, understood, and tracked through each stage of the DISARM process.

Metric	How to measure/track
Phase-wise volume of content or actors identified.	• Track the number of flagged posts and unique actors mapped to each DISARM phase over time using datasets or dashboards.
TTP frequency and evolution tracking	• Measure how often specific TTPs appear and evolve by linking content and actors, then visualising trends through time-series analysis.

TRUSTLAB



Chapter 5

Setting up real-time dashboards to track digital threats on the MediaCloud toolkit

5.Setting up real-time dashboards to track digital threats on the MediaCloud toolkit

5.1. Introduction and purpose

This guide provides a comprehensive overview of setting up dashboards for media monitoring using CivicSignal's MediaCloud, an open-source platform that leverages natural language processing (NLP) to gather and analyse digital news media content across Africa. Adapted from Media Cloud, which was developed at Harvard University's Berkman Klein Centre and the Massachusetts Institute of Technology, this platform enables users to retrieve, visualise, and analyse news stories through continuously updated feeds.

This guide aims to equip users with the knowledge to effectively monitor digital media landscapes and generate data-driven insights through systematic media tracking and dashboard creation.

5.2. Concepts and terminologies

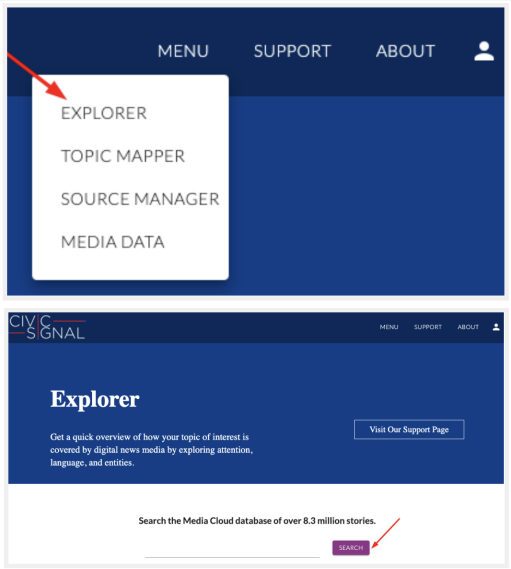
- **Attention:** Measurement of coverage volume through the number of captured stories.
- **Boolean operators:** Search terms (AND, OR, NOT) that refine queries by including or excluding specific parameters.
- **Collection:** Grouping of media sources, typically organised by country (56 country-specific collections available).
- **Explorer:** A tool for searching and analysing how digital news media covers specific topics.
- **Geographic coverage analysis:** Mapping of the spatial distribution of places referenced in stories.
- **Named entity recognition:** Algorithms that extract people and organisations mentioned in articles.
- **Source:** Any organisation, such as fact-checking group, media lab, or newsroom that publishes digital news on CivicSignal.
- **Word Cloud:** Visual representation of term frequency, where text size reflects frequency of mentions.

5.3. Processes and workflows

5.3.1 Best practices

Initial setup

- Log in or register to access Explorer (if you have not registered before, first register as a user).
- Navigate to Menu >> Explorer and click SEARCH.



Screenshot of how to access the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

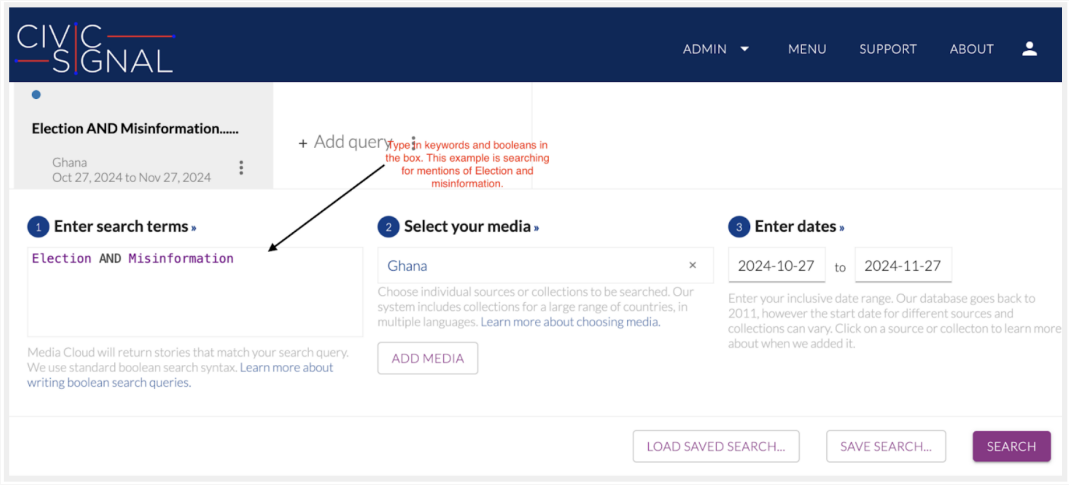
b. Writing effective queries on MediaCloud

- Writing effective queries on MediaCloud
- In the search bar under 'Enter search terms', put the keywords you are interested in. CivicSignal MediaCloud supports Boolean operators to refine searches:

- **AND:** Requires both terms to appear (narrows results).
- **OR:** Includes results with either term (broadens results).
- **NOT:** Excludes specific terms.
- **Parentheses ():** Group terms to control
- query logic.

Here are some examples:

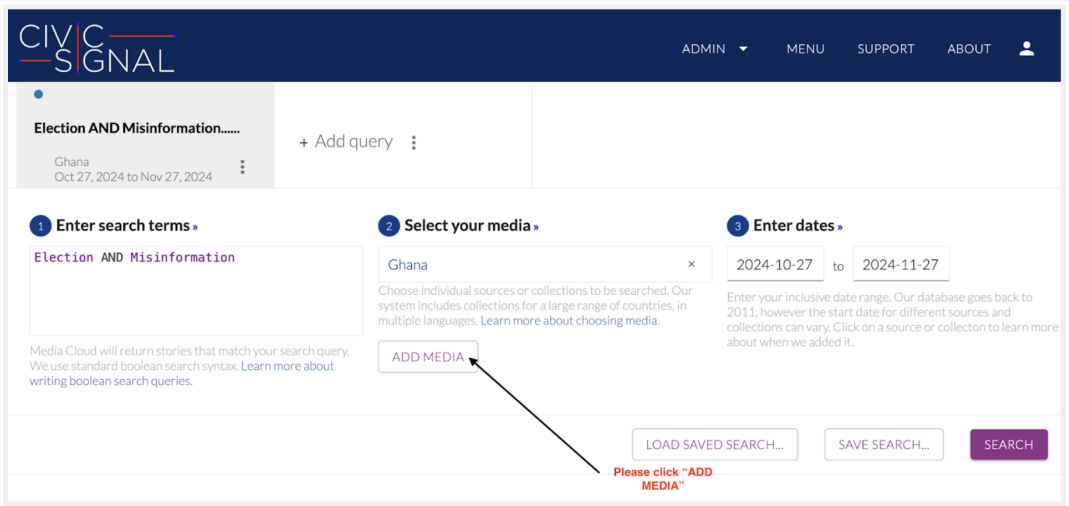
- election OR vote OR ballot → Finds stories with any of these terms.
- election AND misinformation → Finds stories mentioning both terms.
- (election AND misinformation) NOT ('social media' OR 'Facebook') → Excludes stories mentioning social media or Facebook.



Screenshot showing the process of entering search queries in the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

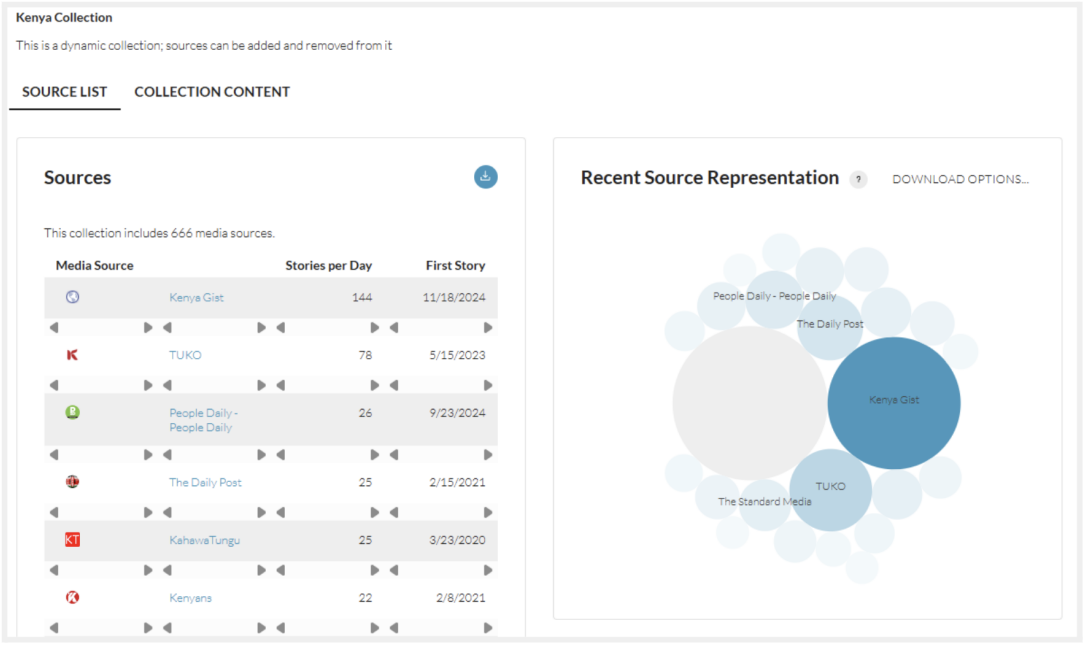
c. Selecting a media source

MediaCloud queries can target multiple collections, a single collection, or a specific media source. To set the search scope, click **'Add media'** in the **'Select your media'** section



Screenshot illustrating the process of selecting media in the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

Use the Source Manager to view the full list of sources available in your country. You can get here by **clicking Menu -> Source Manager -> Browse collections.**



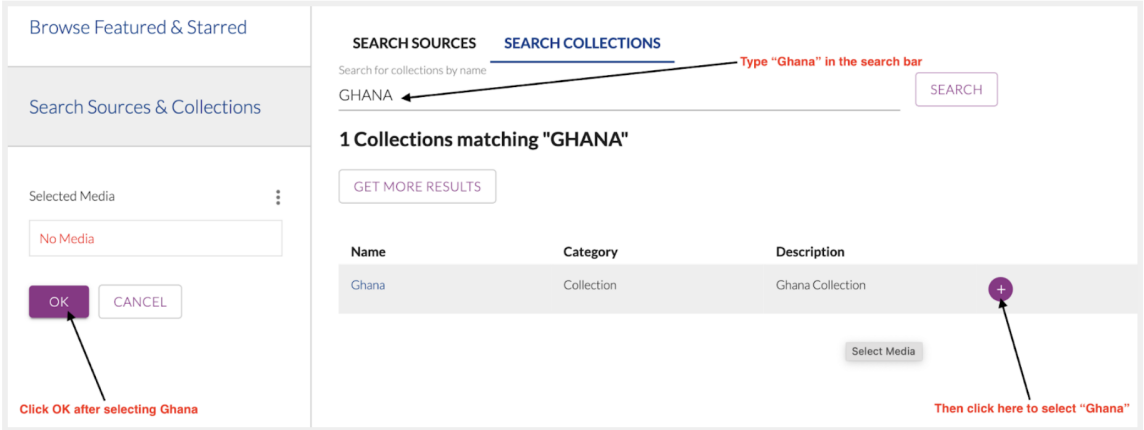
Screenshots of MediaCloud's source manager page for Kenya (Source: CfA's Civic Signal MediaCloud)

You can select media in two ways:

a. Search sources and collections

- Click **'Search Sources & Collections'.**
- Type the collection name (e.g., 'Ghana') in the search bar.
- Click **'SEARCH'**.
- Click the '+' sign to select the collection.
- Click **'OK'**.

To search articles published by a specific source, select **'SEARCH SOURCES'** instead of **'SEARCH COLLECTIONS'** and enter the source name.

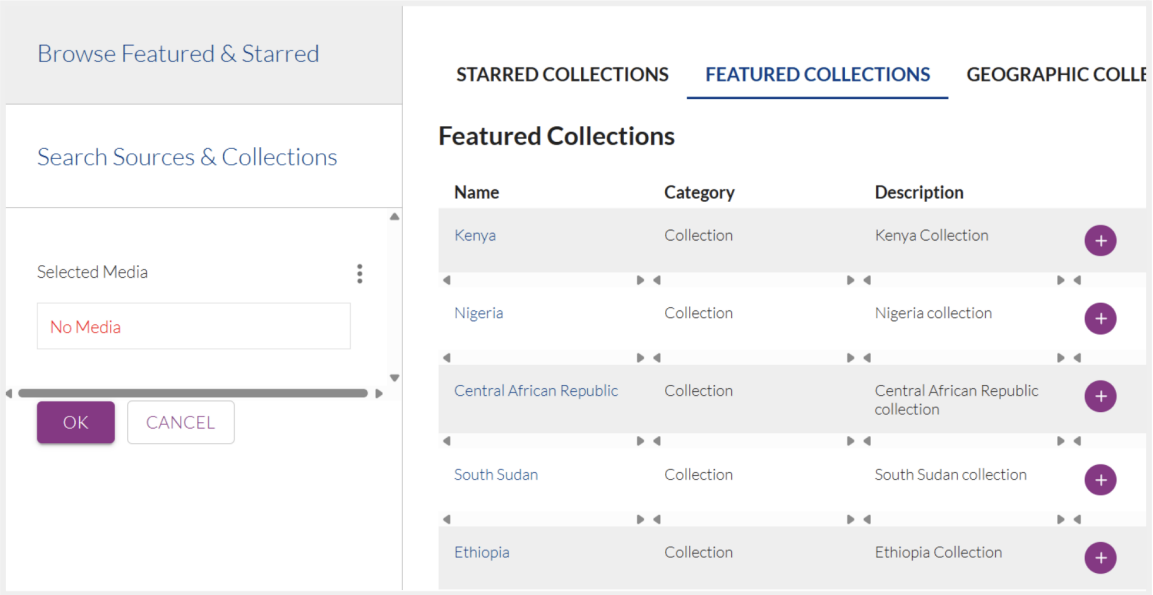


Screenshot illustrating the process of selecting media in the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

b. Using 'Browse Featured & Starred'

Navigate to **'Browse Featured & Starred'.**

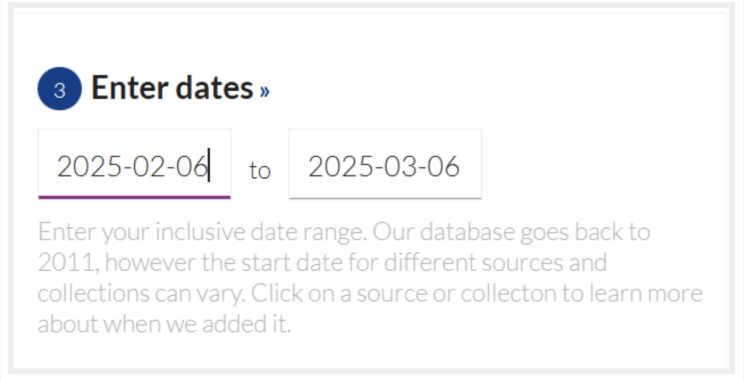
Locate and select the desired source from the list.



Screenshot illustrating the process of browsing featured collections on MediaCloud (Source: CfA's CivicSignal MediaCloud)

Specify the period of interest

- Enter dates in **YYYY-MM-DD** format.
- Once your query is set, click the **'SEARCH'** button to search the database.



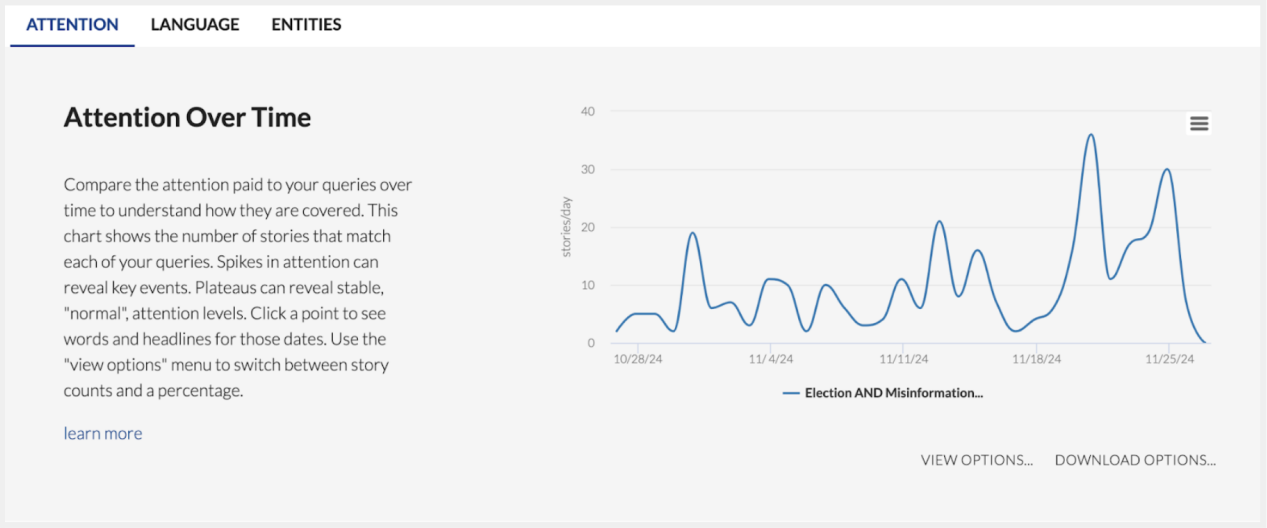
Screenshot illustrating the process of selecting media in the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

5.3.2 Query result analysis

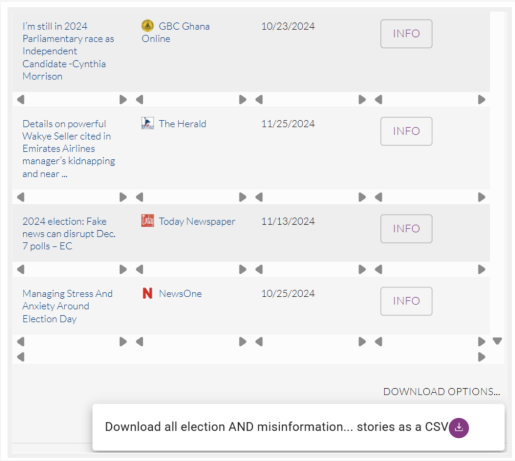
MediaCloud provides three primary analysis dashboards accessible through dedicated tabs: attention, language, and entities. These tools visualise the information from the query result articles.

a. Attention tab

The attention over time graph pinpoints peaks and declines in keyword mentions. Depending on their search needs, users can toggle between viewing the number or percentage of stories that include specific keywords relative to overall news coverage. This dashboard shows when the topic is discussed the most.



Screenshot of MediaCloud's attention tab (Source: CfA's MediaCloud)



You can click on the peak points of this line graph to see a sample of stories published on the day and discussion trends.

MediaCloud allows users to download search query results, including story URLs, in CSV format for further analysis.

Screenshot showing how to download query results from MediaCloud (Source: CfA's MediaCloud)

b. Language tab

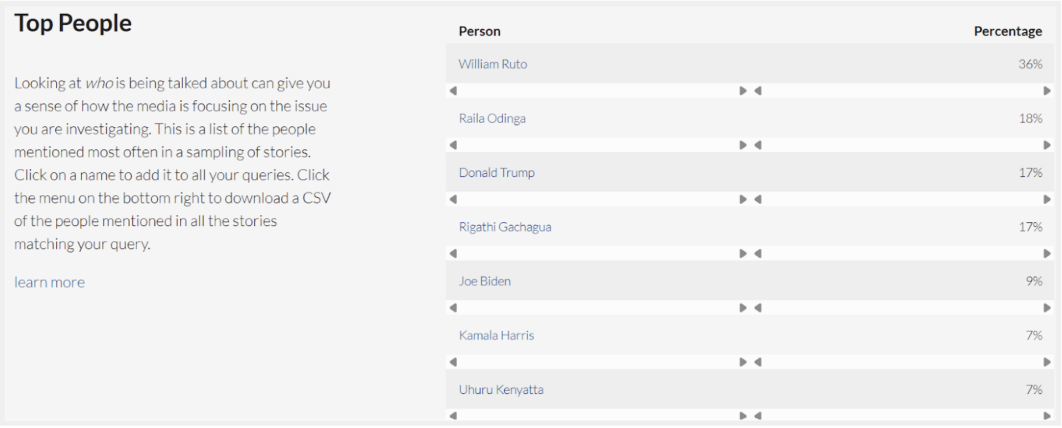
MediaCloud visualises language patterns in search results using word clouds under the 'LANGUAGE' tab. These visualisations highlight the most frequently mentioned terms in articles matching your query. In the 'Top Words' word cloud, text size reflects frequency – larger words appear more often.

The system generates these visualisations from a representative sample of 1,000 stories by default, but users can expand the sample size to 10,000 for a more comprehensive analysis, though this alteration requires additional processing time.

Screenshot illustrating the process of selecting media in the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

c. Entity tab

CivicSignal's MediaCloud highlights the top people and organisations mentioned in stories under the 'ENTITY' tab.



Screenshots of MediaCloud's source manager page for Kenya (Source: CfA's Civic Signal MediaCloud)

MediaCloud uses named entity recognition algorithms to extract and rank people and organisations mentioned in articles based on their frequency of appearance. The system calculates the percentage of articles containing each entity, helping researchers identify the most prominent actors in a given topic.

Geographic coverage analysis maps the spatial distribution of places referenced in stories, revealing a publication's reporting focus beyond its home country. For instance, Nigerian outlet 'The Punch' reports on domestic events and international news from locations such as the US. However, geographic tags appear only when the system successfully extracts and recognises location data.

5.4. Roles and responsibilities

Media analyst: Formulates effective search queries, interprets attention patterns, and generates reports on media coverage trends. The media analyst should be familiar with the media landscape of the country of interest or the topic of concern so they can identify the most relevant keywords and actors to track.

The person should also be proficient with Google Sheets to clean downloaded data in spreadsheet format if needed. Having an additional data analysis background would be beneficial for deeper insights and more sophisticated reporting.

5.5. Tools and resources

Table 9; below highlights the tools and resources required to track digital threats on the MediaCloud toolkit.

Tool	Purpose
MediaCloud Source Manager	• Review sources available for each country.
MediaCloud Explorer	• Searching and analysing digital news media coverage on specific topics.
CSV Export Function	• Downloading search query results, including story URLs, for further analysis.

5.6. Best practices

Query construction:

i. Use specific keywords rather than general terms to narrow results.
ii. Test multiple query variations to ensure comprehensive coverage.

b. Documentation and analysis:

i. Share queries and methodologies with team members for consistency.
ii. Document analysis approaches for future reference and replication.
iii. Combine quantitative media data with qualitative content analysis.

5.7. Metrics and performance indicators

Track the essential metrics outlined in table 10 below to evaluate the effectiveness of your dashboard:

Tool	Purpose
MediaCloud Source Manager	The percentage of relevant stories captured per search query and the percentage of results matching monitoring objectives versus irrelevant results. Queries may need to be refined to exclude irrelevant results.
CSV Export Function	The percentage of results leading to further investigation.

TRUSTLAB

Chapter 6
Building lexicons and watchlists

6.Building lexicon and watchlists

6.1 Introduction and purpose

This section outlines the objectives and scope of monitoring and listening systems. It explains how these systems support the consistent execution of digital threat detection, focusing on capturing emerging trends, misinformation campaigns, and digital threats. The goal is to help users understand how monitoring tools function to track online narratives and enhance the analysis of harmful content.

Importance of monitoring/listening systems

- **Capture unseen posts:** Even with expertise in social media investigations, spotting every instance of misinformation within your region is impossible. Monitoring systems, such as Google Alerts, help you capture posts you might otherwise miss by sending email alerts when keywords such as 'beware', 'fake', or 'scam' are used.
- **Diversification:** Focusing on a specific area may limit you to a particular trend, such as hoaxes. Monitoring systems allow for broader coverage, enabling you to track misinformation on various topics. For instance, using a Google form or WhatsApp tipline, you can ask your audience to report potential misinformation, resulting in a diverse range of claims for debunking. Here is how to create a Google form.
- **Filtering information:** Manual review is impractical with millions of daily social media posts. Monitoring systems filter data to match your search criteria. Features such as advanced X search allow you to filter posts by account, language, and timeframe, ensuring a focused and relevant dataset.
- **Saves time:** Monitoring systems streamline the process of identifying claims. Using Facebook or X lists, you can group accounts known for spreading misinformation and quickly check their posts without scrolling through individual timelines. Steps to create Facebook and X lists are outlined here and here, respectively.

6.2. Concepts and terminologies

Lexicons are structured databases of keywords, phrases, and thematic terms used to track narratives, trends, and entities in online spaces. They play a crucial role in monitoring disinformation, electoral discourse, and hate speech by identifying 'trigger' terms that incite audiences.

Creating a lexicon involves selecting highly relevant terms aligned with research objectives, whether tracking coordinated campaigns, elections, or specific hashtags. A well-compiled lexicon enhances targeted analysis, ensuring precise interpretation of online conversations.

Lexicons evolve with digital discourse, adapting to emerging terminologies and patterns to maintain their effectiveness in media monitoring and investigative work.

Table 9; below highlights the tools and resources required to track digital threats on the MediaCloud toolkit.

Classifier	Description
Keywords	These are context-specific words or phrases that carry particular significance in digital conversations.
Terms	Relevant keywords or phrases that actors often weaponise to attack, demean, or insult groups or individuals. These include derogatory terms, slurs, and coded language that dehumanise, incite hostility, or reinforce stereotypes. Their impact varies based on cultural and contextual factors but is consistently aimed at escalating conflict and division.
Variant terms	Synonyms or variations of terms that carry the same meaning but are used in different contexts, dialects, or pronunciations.
Language	Refers to the medium of communication, encompassing spoken, written, and digital expressions in languages such as Arabic, English, Igbo, Kiswahili, and Shona.
Offensiveness	A severity scale assigned to words and phrases based on their level of unpleasantness, ranging from mild insults to highly aggressive slurs. The impact of such language varies depending on cultural context, intent, and the targeted group, helping to assess the degree of harm or inflammatory potential in communication.
Regions	Refers to geographical areas associated with specific ethnic languages or communities. These locations act as cultural and linguistic hubs, shaping contexts, dialects, social norms, and traditions.
Context	Refers to the circumstances surrounding a message, shaping its meaning and impact. Factors such as ancestry, class, ethnicity, extremism, gender, nationality, race, religion, slurs, violence, and xenophobia influence how a message is perceived. These elements determine tone, intent, and the potential for escalation or harm in communication.

The data schema for classifying lexicons provides a structured approach to organising and analysing language patterns. It includes factors to consider when building lexicons, including:

a. **Context:** At its core, the schema incorporates an index column, which assigns a unique identifier to each entry, allowing for systematic tracking and reference. Additionally, a category column serves as the foundation for classification, helping contextualise terms based on their relevance to different socio-political themes.

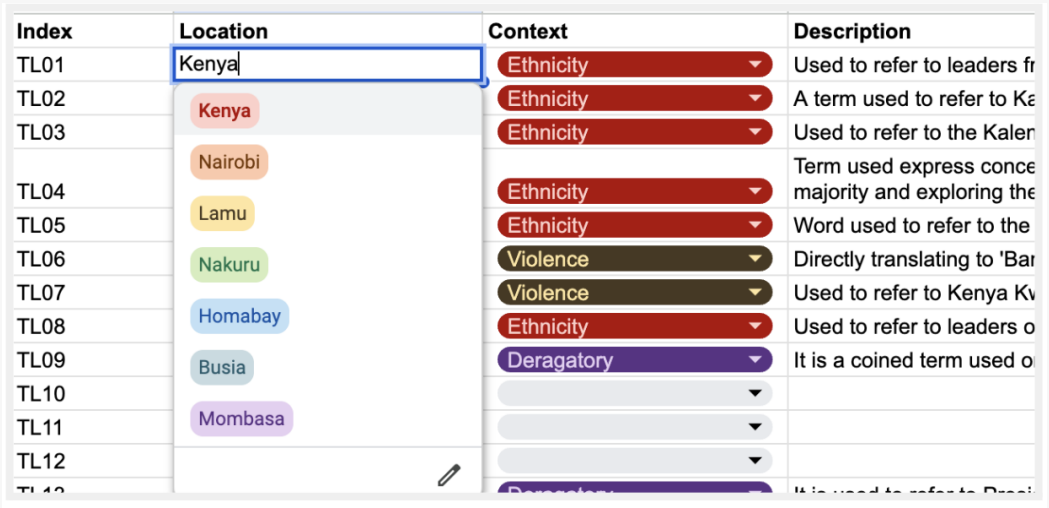
These categories are not exhaustive and may shift depending on geographical and cultural contexts.
Classifications include class, ethnicity, ethnic slurs, and religion, which capture language reflecting racial discrimination, religious bias, or social hierarchies. Colour, nationality, and race shape discourse on identity, inclusion, and exclusion, while ancestry can highlight heritage or be weaponised for exclusion.
Gender-based language reveals societal attitudes, including gendered insults, misogyny, and patriarchy. Extremism and xenophobia propagate hate, ideological bias, or radicalisation, while violence-related terms highlight rhetoric inciting harm or conflict.

Index	Category	Description
cat_01	Class	Keywords based on classes in society, each with their own unique set of values and norms. The upper class is typically associated with wealth and privilege, while the lower class is often associated with poverty and hardship. The use of these keywords is to to promote division or hatred between the social and economic classes
cat_02	Religion	This includes keywords that target people of other religions, with an aim to undermine, hate and promote division between religions and within religions.
cat_03	Ethnicity/ Ethnic Slur	Keywords that are used to target people of other races with an aim to promote hate, division or undermine.
cat_04	Nationality	Keywords that are used to target people of other nationalities with an aim to promote hate, division or undermine. This can be the same race but different nationalities.
cat_05	Race	Keywords that are used to target people of other races with an aim to promote hate, division or undermine.
cat_06	Colour	Keywords that are used to target people of different skin color with an aim to promote hate, to show non-belonging and inferiority inorder. In most cases, this is perpetrated by people of same race but discriminate based on skin complexion (colorism with hate).
cat_07	Ancestry	Keywords based on ancestry such as clans, which when used they promote hate towards a particular clan in the same ethnicity group.
cat_08	Gender	keywords that are targeted towards a particular gender with an aim to undermine and promote hatred towards that particular gender. This can be both perpetrated by different genders i.e misogyny, or same gender i.e internalized misogyny.
cat_09	Xenophobia	This includes keywords that show prejudice against people from other countries, they can be same race, same ethnicity but different nationalities.
cat_10	Extremism	keywords that promote extremism or fundamentalism that promote hatred, division and undermining. This key words intersect with both political views and religion views that are extreme/ fundamentalism.
cat_11	Violence	Keywords that can be used in combination with actor watchlist to incite the public into violence

Screengrab of an example of a context category to guide in lexicon identification (Source: CfA using TrustLab lexicon data)

These categories are not exhaustive and may shift depending on geographical and cultural contexts.
Classifications include class, ethnicity, ethnic slurs, and religion, which capture language reflecting racial discrimination, religious bias, or social hierarchies. Colour, nationality, and race shape discourse on identity, inclusion, and exclusion, while ancestry can highlight heritage or be weaponised for exclusion.
Gender-based language reveals societal attitudes, including gendered insults, misogyny, and patriarchy. Extremism and xenophobia propagate hate, ideological bias, or radicalisation, while violence-related terms highlight rhetoric inciting harm or conflict.

b. Location



Screengrab of an example showcasing the mapping of Kenyan counties (Source: CfA using the TrustLab lexicon database)

Geographical classification of lexicons helps capture regional language nuances, reducing false positives and improving the identification of harmful content. Words may have different meanings across regions, making context crucial in distinguishing between neutral and harmful usage. For example, in Kenya, lexicons were categorised by focus counties. Assigning an index number based on geography allowed for a more precise classification of ethnic terms, ensuring accurate contextual analysis.

6.2.1 Key term classification

Key term classification is essential when building a lexicon database. Terms are categorised as phrases or words to refine monitoring and query development. Offensive classification assigns a scale to gauge derogatory or harmful usage. Language and location categorisation ensure recognition of regional variations. Context is crucial in assessing offensiveness, especially during elections on platforms such as Facebook, Instagram, TikTok, and X.

Example of key term classification in the Kenyan context:

i. **Keyword:** 'Madoadoa' (Kiswahili word for 'blemishes').
ii. **Classification:** Highly offensive, used nationally by some politicians.
iii. **Context:** Used to incite violence or promote ethnic cleansing by targeting individuals perceived as outsiders or non-native to a particular region or political jurisdiction. In political discourse, it polarises communities.
iv. **Classification in Kenya:** Considered a hate lexicon due to its potential to incite violence and promote division.

b. Location

Madoadoa		Word	Highly offensive
Madude		Word	Extrememly offensive
Mageryenge		Word	**Highly offensive**
Mende		Word	Mildly offensive
Mombasa ni yetu wabaara wa		Phrase	Moderately offensive
Muhajir		Word	

Screenshot of the classification of the hate term 'madoadoa', using a scale to show its level of offensiveness (Source: CfA using the Kenya lexicon database)

6.2.2 Lexicon building based on identity factors

Lexicon building involves systematically compiling terms and phrases that reflect identity categories such as ethnicity, nationality, race, region, and sex. These distinctions help capture the nuances of language use across different social and cultural settings, shaping communication patterns, social perceptions, and group identities.

Term	Variant	Type	Offensiveness	Language	location	I2 ntext	Des
Kimurkeldet	Kimurkelda	Word	Highly offensive	Kalenjin	Rift valley	Ethnicity/ Ethnic Slur	This
Kura haramu, kura ukafiri		Phrase	Highly offensive	Swahili	Coastal	Class	
Kwekwe	makwekwe	Word	Highly offensive	Swahili	Coastal	Religion	
Lazima tushinde aidha kupitia		Phrase	Mildly offensive	Swahili	Kenya	**Ethnicity/ Ethnic Slur**	
Madoadoa		Word	Highly offensive	Swahili	Kenya	Nationality	
Madude		Word	Mildly offensive	Swahili	Coastal	Race	
Mageryenge		Word	Mildly offensive	Mijikenda	Coastal	Colour	
Mende		Word	Mildly offensive	Swahili	Coastal	Ancestry	
Mombasa ni yetu wabaara wa		Phrase	Mildly offensive	Swahili	Coastal	Gender	
Muhajir		Word	Mildly offensive	Kenyan Arabs	Coastal	Xenophobia	
Munafikin	munafik	Word	Mildly offensive	Swahili	Coastal	Extremism	
Muslims are al-Shabaab		Phrase	Highly offensive	English	Kenya	Violence	
Ngawira		Word	Mildly offensive	Swahili	Coastal		
Nitakutoa supu (kutoa damu); Ntakutoa Supu, kunywa damu ya mtu);	Hatuwawachi hadi tumwage damu	Phrase	Extrememly offensive	Swahili	Kenya		

An overview of the lexicon schema with the various classifications (Source: CfA using the Kenya lexicon database)

Below is a breakdown of major terms related to identity and classification, including ethnicity, nationality, race, region, and sex. These terms help to understand the various ways people are categorised and how these categories intersect in different social, political, and cultural contexts.

- **Gender:** Words and phrases related to sex encompass gender-specific language, pronouns, and terms reflecting identities, social roles, or stereotypes. A lexicon in this category helps identify discrimination or misogyny, monitor gender portrayals, and track gendered language patterns.

- **Ethnicity:** Ethnic-specific lexicons encompass terms tied to ethnic groups, identities, and traditions. They help track culturally significant phrases, divisive language, and ethnic slurs. For instance, negative stereotypes linking communities to traits such as laziness or witchcraft reinforce harmful biases.

- **Regional:** Regional lexicons track terms unique to specific areas, including slang and dialects. They help identify location-based narratives and variations in meaning across regions. For example, phrases common in West Africa may have different connotations in East Africa, influencing interpretation.

- **Nationality:** Nationality-based lexicons capture terms linked to national identity, stereotypes, and xenophobia. They help track references to countries, national figures, and symbols, especially when used to incite hostility or reinforce bias.

- **Race:** Race-related lexicons identify terms linked to racial identity, racialised language, and slurs. They are essential for analysing racial discourse, monitoring hate speech, and tracking racist rhetoric.

6.2.3 Lexicon building based on identity factors

Event-based lexicons are specialised collections of phrases, terminologies, and words tied to specific events, occurrences, or time-bound situations. These terms capture the language and narrative patterns surrounding events, making them essential for social media monitoring and analysis. They help track and interpret how public discourse shapes responses to elections, major incidents, protests, social movements, and other impactful occurrences.

Considerations for event-based lexicon building:

- **Identifying event-specific keywords:** Terms linked to particular events, such as the names of political gatherings, protest slogans, or trending hashtags, are central to this type of lexicon. These keywords help monitor how events are framed, perceived, and reported in real time.
- **Relevance:** Some phrases are only relevant during the event itself, while others evolve in meaning as the situation unfolds. Event-based lexicons must be continuously updated to reflect these changes. For example, the hashtag #EndSARS gained significance during the Nigerian anti-government protests that took place between 01 and 10 August 2024, but later took on broader connotations in discussions on police brutality.
- **Emotion and sentiment:** Events often generate emotionally charged language. Lexicons should include terms that capture a range of sentiments, such as outrage or calls to action (e.g., burn houses), as these significantly influence public opinion and narrative framing.

- **Monitoring the evolution of events:** The language surrounding an event can shift rapidly. A peaceful protest may escalate into violent clashes, changing the terms used to describe it. An effective event-based lexicon must be adaptable to these shifts. For instance, in Kenya, the #RejectFinanceBill2024 movement began as a peaceful anti-government outcry on social media in May 2024.

  However, by 18 June 2024, the protests, primarily led by Generation Z, escalated into nationwide demonstrations that turned violent due to confrontations with police over a month-long period.

  Counter-narratives emerged as the protests gained momentum, shifting the discourse to anti-Gen Z sentiments. Opponents sought to discredit the movement by associating Gen Z protesters with the lesbian, gay, bisexual, and transgender (LGBTQ+) community, using derogatory hashtags such as #GayZ to frame the demonstrators as being members of the LGBTQ+ community.

6.2.4 Hate lexicon

A hate lexicon is a curated collection of coded language, keywords, phrases, and slurs commonly used in hate speech. It serves as a critical tool for identifying and categorising harmful language in both online and offline spaces.

The UN defines hate speech as any form of communication – behavioural, spoken, or written – that discriminates against or attacks individuals or groups based on identity factors such as ethnicity, colour, descent, gender, nationality, race, or religion. There is no universal definition of hate speech under international human rights law, as it remains contested in relation to equality, freedom of expression, and non-discrimination.

Leveraging expertise from specialised organisations is essential to building an effective hate lexicon. These organisations focus on identifying hateful words, slang, and coded language across different regions, making them valuable resources in lexicon development. Some initiatives include: Dangerous Speech Project, Hatebase, Masakhane Project, and PeaceTech Lab.

Components of a hate lexicon

a. **Hate speech terms:** These are explicit words or slurs used to attack, demean, or insult individuals or communities. They may include anti-LGBTQ+ phrases that are offensive, misogynistic terms, or racial slurs.

	Index	Term	Variant	Type	Offensiveness	Language	location	Context
1	Ter_01	Kimurkeldet	Kimurkelda	Word	Highly offensive	Kalenjin	Rift valley	Ethnicity/ Ethnic S
2	Ter_02	Kura haramu, kura ukafiri		Phrase	Highly offensive	Swahili	Coastal	Religion
3	Ter_03	Kwekwe	makwekwe	Word	Highly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
4	Ter_04	Lazima tushinde aidha kupitia		Phrase	Mildly offensive	Swahili	Kenya	Extremism
5	Ter_05	Madoadoa		Word	Mildly offensive	Swahili	Kenya	Ethnicity/ Ethnic S
6	Ter_06	Madude		Word	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
7	Ter_07	Mageryenge		Word	Mildly offensive	Mijikenda	Coastal	Ethnicity/ Ethnic S
8	Ter_08	Mende		Word	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
9	Ter_09	Mombasa ni yetu wabaara wa		Phrase	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
10	Ter_10	Muhajir		Word	Mildly offensive	Kenyan Arabs	Coastal	Religion
11	Ter_11	Munafikin	munafik	Word	Mildly offensive	Swahili	Coastal	Religion
12	Ter_12	Muslims are al-Shabaab		Phrase	Highly offensive	English	Kenya	Religion

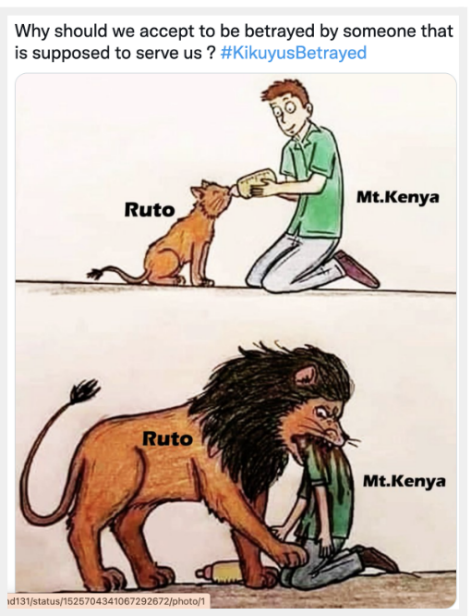
Screenshot showing a collection of Kenyan hate terms with various classifications (Source: CfA using the Kenya lexicon database)

b. Coded language: Malign actors often use coded language to avoid detection. This can include seemingly innocent words with hidden meanings or terms that have been reappropriated to convey hate within specific communities.

Screenshot highlighting the use of coded language referring to a certain ethnic group in Kenya (Source:CfA via X)

c. Use of imagery: Hate speech is not always limited to words. It can also include specific phrases, slogans, or symbols that convey hostile intent. Capturing these expressions and visual elements (e.g., hate symbols or imagery) expands the lexicon's capability to detect forms of hateful content.

A dynamic hate lexicon must be continuously updated to track these shifts, ensuring real-time identification and mitigation of harmful content.



Screenshot showing the use of imagery to spread hateful content during the 2022 Kenyan elections (Source: CfA via ADDO)

6.2.5 Freedom of expression vs hate speech

The International Covenant on Civil and Political Rights defines freedom of expression as the right to hold opinions without interference, including the right to seek, receive, and share information across borders through any media.

The Rabat Plan of Action sets a high threshold for restricting freedom of expression, particularly regarding incitement to hatred. It applies a six-part test to assess whether speech warrants limitation: context, content and form, extent of dissemination, intent, likelihood and imminence of harm, and speaker.

(1) **Context**: Context is of great importance when assessing whether particular statements are likely to incite discrimination, hostility or violence against the target group, and it may have a direct bearing on both intent and/or causation. Analysis of the context should place the speech act within the social and political context prevalent at the time the speech was made and disseminated;

(2) **Speaker**: The speaker's position or status in the society should be considered, specifically the individual's or organization's standing in the context of the audience to whom the speech is directed;

(3) **Intent**: Article 20 of the ICCPR anticipates intent. Negligence and recklessness are not sufficient for an act to be an offence under article 20 of the ICCPR, as this article provides for "advocacy" and "incitement" rather than the mere distribution or circulation of material. In this regard, it requires the activation of a triangular relationship between the object and subject of the speech act as well as the audience;

(4) **Content and form**: The content of the speech constitutes one of the key foci of the court's deliberations and is a critical element of incitement. Content analysis may include the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed in the speech or the balance struck between arguments deployed;

(5) **Extent of the speech act**: Extent includes such elements as the reach of the speech act, its public nature, its magnitude and size of its audience. Other elements to consider include whether the speech is public, what means of dissemination are used, for example by a single leaflet or broadcast in the mainstream media or via the Internet, the frequency, the quantity and the extent of the communications, whether the audience had the means to act on the incitement, whether the statement (or work) is circulated in a restricted environment or widely accessible to the general public; and

(6) **Likelihood, including imminence**: Incitement, by definition, is an inchoate crime. The action advocated through incitement speech does not have to be committed for said speech to amount to a crime. Nevertheless, some degree of risk of harm must be identified. It means that the courts will have to determine that there was a reasonable probability that the speech would succeed in inciting actual action against the target group, recognizing that such causation should be rather direct.

The Rabat Plan of Action's six-part threshold test for restricting freedom of expression in detail (Source: CfA via the Rabat Plan of Action)

6.3. Processes and workflows

Types of monitoring and listening systems

Some of the listening and monitoring systems for lexicons include:

- Creating Facebook and X lists.
- Following multiple social and political Facebook groups, especially those with large followings.
- Monitoring the trending X topics.
- Setting up Google Alerts.
- Joining several WhatsApp groups.
- Crowdsourcing, which involves asking people to share any suspicious information they see online through an email, tipline, web form, or social media platforms.

The process of building a watchlist includes:

- Use tools such as Google Alerts, social media lists, and WhatsApp groups to track information and set up monitoring systems.
- Continuously monitor social platforms for emerging narratives and potential misinformation campaigns to identify key actors and threats.
- Engage the community to report suspicious content through various channels, such as tiplines or web forms.
- Develop a structured lexicon to track harmful or misleading terms. Regularly update the lexicon to adapt to evolving discourse.
- Use monitoring systems to filter and categorise posts based on defined criteria such as keywords, sentiment, and offensiveness.

6.4. Roles and responsibilities

Set up and maintain monitoring databases or watchlists, ensuring regular lexicon updates and tracking emerging trends. Additionally, work with lexicons to categorise and assess the impact of identified threats.

6.5. Tools and resources

Table 12; below highlights the tools and resources required to build lexicons and watchlists.

Tool	Purpose
Google Alerts	• Monitors online content based on specified keywords.
Social media lists	• Used to track specific accounts or topics on platforms such as Facebook or X.
Lexicon databases (e.g PeaceTech Lab)	• Structured terms and phrases that track harmful narratives.

6.6. Best practices

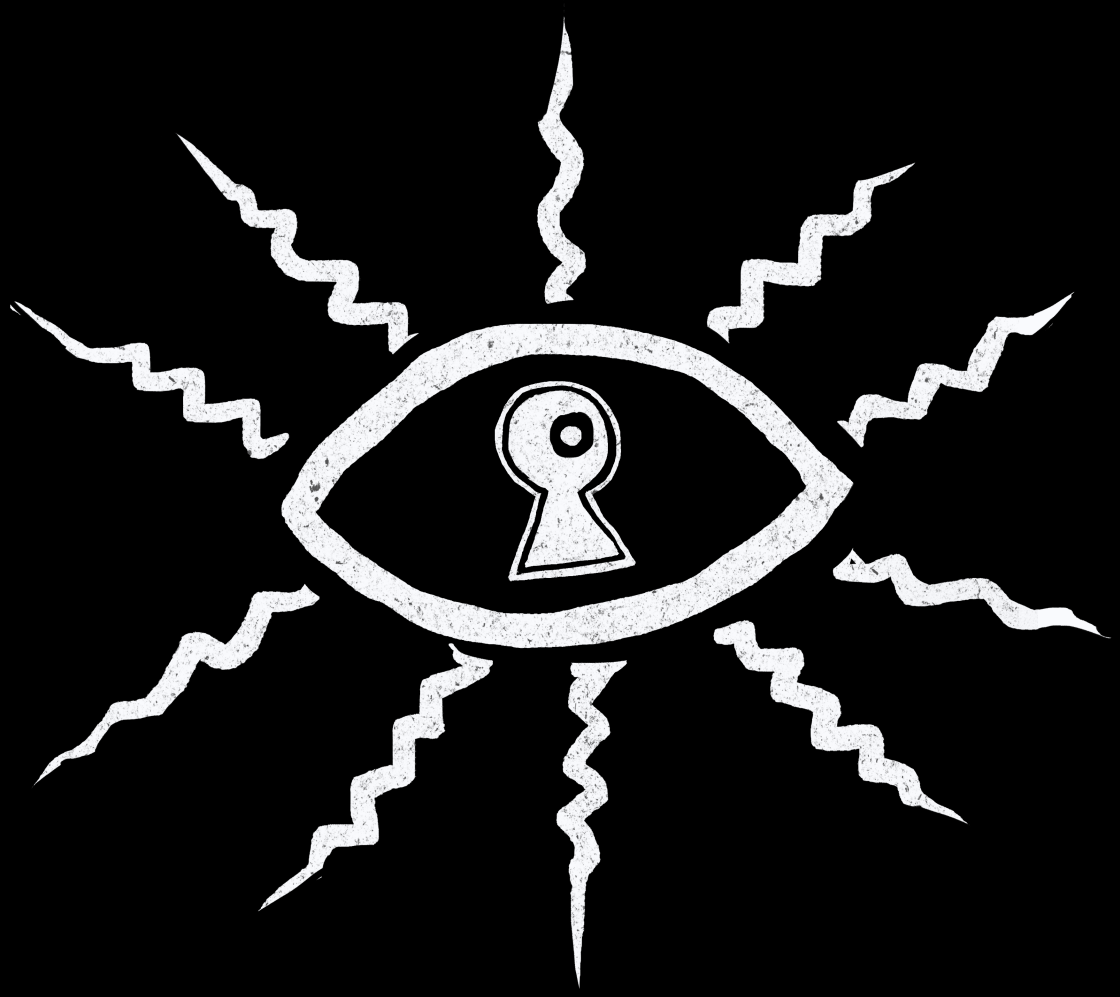
- **Diversification:** Ensure monitoring covers a wide range of topics to identify various forms of misinformation beyond a narrow focus on hoaxes.
- **Data filtering:** Use advanced search filters to reduce noise and focus on relevant posts.
- **Community engagement:** Foster an active relationship with users, encouraging them to report suspicious content via email or social media platforms.
- **Regular updates:** Continuously update lexicons to reflect new trends, regional variations, and evolving language.

6.7. Metrics and performance indicators

To track progress and outcomes, use the metrics outlined in **table 13** below:

Metric	How to measure/track
Engagement metrics	• Measure user involvement in crowdsourcing efforts, including the number of reports submitted.
System accuracy	• Monitor the effectiveness of lexicon filtering in identifying relevant content.
Reach and impact	• Track the number of posts identified and flagged for review, as well as subsequent actions taken (e.g., content removal or fact-checking).

TRUSTLAB



Chapter 7
Tips for producing threat reports

7.Tips for Producing threat reports

7.1 Introduction and purpose

This section aims to provide a structured framework for writing a threat report. The purpose of the threat report is to assist CBOs/ CSOs or partners in identifying, analysing, and responding to emerging disinformation incidents by providing timely, evidence-based insights.

The report aims to highlight key actors, tactics, and narratives associated with the campaign, assess its virality, and pinpoint the phase of the attack using recognised frameworks like the DISARM. By outlining clear processes and roles, this guide ensures that threats are identified early, documented comprehensively, and addressed effectively, supporting consistent execution and ongoing improvements in counteraction strategies.

7.2. Concepts and terminologies

- **Actors:** Individuals or groups behind disinformation campaigns, ranging from state actors to coordinated fake accounts.
- **Archiving:** Storing evidence of disinformation activities for forensic analysis and future reference.
- **Coordinated information manipulation:** Organised efforts to manipulate public opinion through misinformation, often involving multiple actors.
- **Kill chain phases:** A model that divides operations into phases (preparation, execution, post-execution) to assess intervention points.
- **Narratives:** Influence operations or disinformation campaigns used to influence perceptions or agendas.
- **Network analysis:** Mapping the interactions between actors, content, and platforms to understand how disinformation spreads.
- **Virality check:** A method to assess the scale of a disinformation incident based on its speed and reach across platforms.

7.3. Processes and workflows

Step 1: Initial identification

Monitor platforms such as Facebook, Telegram, TikTok, or X. Use free or paid tools (e.g., Meltwater, Talkwalker Alerts, X Pro (formerly TweetDeck), OR X search function) to track keywords, hashtags, trends, and influential accounts.

How to tell if a finding is important:

- Use the Breakout Scale (a tool or scale to measure virality or momentum) to check how fast the narrative or post is spreading.
- Ask: Is it trending? Are influential or verified accounts sharing it? Is the media picking it up?

Step 2: TTP check

The goal here is to check whether what you are seeing is a known manipulation technique. Compare the activity you have identified with the DISARM framework.

Some common TTPs include:

- Distort facts (T0023).
- Leverage existing narratives (T0003).
- Create dedicated hashtags (T0104.006).

Step 3: Kill chain phase mapping
The goal is to understand where in the disinformation operation the activity fits. Use the DISARM kill chain model or a similar framework to map the behaviour to a phase:

- Preparation, such as building fake accounts or collecting target information.
- Execution, such as posting the content and amplifying narratives.
- Post-execution/evaluation, such as tracking performance or adjusting content.

Step 4: Actor identification
Compare the suspicious accounts or websites to your watchlist (a list of previously identified bad actors or suspicious groups). Note their usernames, bios, links to other accounts, or location clues.

Step 5: Narrative check
See if the message being pushed matches known disinformation themes. Use a lexicon watchlist or keyword list to match the narrative to previously seen harmful themes. Examples of themes include 'election rigging' or 'ethnic supremacy'.

Step 6: Network analysis
The goal here is to check if the activity is coordinated or not. You need to:

- Export social media data (e.g., mentions, replies, or reposts) and upload it to a tool such as Gephi or NodeXL.
- Analyse connections between accounts: Are a group of accounts all posting the same content at the same time?

Step 7: Archiving evidence
Use tools such as perma.cc to archive social media links or websites so you have a permanent snapshot. Organise your data in AirTable, Excel, or Google Sheets to keep track of URLs, posts, usernames, and screenshots.

7.4. Roles and responsibilities

These are the skills required for writing a threat report:

- **Narrative monitoring:** The ability to track and monitor ongoing online narratives across platforms.
- **Threshold assessment:** Evaluating whether narratives meet the criteria for inclusion in strategic or threat reports.
- **Intent analysis:** Competence in reviewing content for signs of malicious intent and identifying indicators of potential amplification or manipulation.
- **Virality assessment:** Proficiency in evaluating narrative spread using tools such as the Breakout Scale.
- **TTP mapping:** Identifying TTPs using the DISARM schema, determining the phase of an incident within the kill chain framework, and providing contextual analysis.
- **Entity verification:** Cross-referencing actors against existing watchlists to detect known threat entities.
- **Narrative classification:** Matching incident narratives to known entries in lexicon databases and updating with new terms or actors.
- **Network analysis:** Using tools such as Gephi to produce network graphs that identify coordinated or suspicious activities.
- **Evidence archiving:** Properly archiving digital evidence using tools such as perma.cc and AirTable for long-term reference and traceability.

7.5. Tools and resources

Table 15; below outlines the tools necessary to produce a threat report.

Tool	Purpose
Breakout Scale	• Used for assessing the virality of an incident and determining its prioritisation.
DISARM framework	• A framework to identify and categorise the tactics used by malicious actors.
Lexicon watchlist/database	• Databases are used for recording and storing harmful content trends.
Gephi	• A platform for generating network analysis graphs to track and assess coordinated influence campaigns.
Perma.cc	• A tool used for archiving and storing incident data and evidence.

7.6. Best practices

- Ensure the threat report is compiled and shared on time to enable appropriate intervention.
- Always verify facts and use reliable sources to check narratives, actors, and TTPs.
- Organise reports logically, with clear sections for each part of the workflow (incident summary, narrative analysis, actor analysis, TTP analysis, and countermeasures).
- Maintain close communication with all the involved teams to prioritise the most relevant and impactful narratives.

7.7. Metrics and performance indicators

The metrics outlined in **table 16** below are used to assess the quality and impact of a threat report, ensuring that disinformation incidents are properly prioritised and analysed.

Metric	How to measure/track
Incident tracking	• Track how often incidents meet the virality threshold for prioritisation and the types of incidents that require the most attention.
TTP identification and mapping	• Measure how accurately TTPs are identified and mapped to known patterns.