COMMUNITY INFORMATION RESILIENCE PLATFORM

CIRP Intelligence Portal

Platform Workflow & Capabilities Guide

Multi-agency intelligence dashboard for monitoring ethnic and diaspora media narratives across 6+ languages and 50+ sources

51+ **Sources**	**6+** **Languages**	**35** **Communities**	**5** **Threat Categories**	**4** **Alert Tiers**

CLASSIFICATION: UNCLASSIFIED // LAW ENFORCEMENT SENSITIVE (U/LES) | Version 2.0 | May 02, 2026

This document is intended for authorized government and community partner personnel only.

Table of Contents

1. Executive Summary

The Community Information Resilience Platform (CIRP) is a multi-agency intelligence dashboard designed to monitor, classify, and respond to harmful narratives circulating in ethnic and diaspora media ecosystems across the United States. Built to serve government analysts, community liaison officers, and authorized partner organizations, CIRP provides real-time threat detection across 6+ languages and 50+ ethnic media sources — including Spanish, Arabic, Hindi, Haitian Creole, Russian, Somali, Mandarin, and French outlets.

Disinformation, criminal exploitation narratives, health misinformation, and coordinated foreign influence operations increasingly target immigrant and diaspora communities through ethnic media channels that are often unmonitored by traditional intelligence infrastructure. CIRP closes that gap by combining automated web scraping, natural language processing (NLP), machine learning classification, and AI-generated counter-narrative drafting — all surfaced in a secure, role-gated portal.

Key Platform Objectives

- Provide early warning of harmful narratives targeting diaspora communities before they spread.
- Classify threats by type, severity, and targeted community using a hybrid keyword + ML engine.
- Automatically route high-priority alerts to human reviewers before multi-agency deployment.
- Generate culturally tailored AI counter-narratives to assist response teams.
- Maintain a complete audit trail of all platform activity for oversight and compliance.
- Support 5 distinct user tiers — from community partners to government administrators.

2. System Architecture Overview

CIRP is a Python-based web application served by a Gunicorn WSGI server (4 workers, 2 threads each) on port 5000. The platform is composed of five integrated subsystems that operate in a coordinated pipeline:

Subsystem	Technology	Role
Web Scraper	Python / BeautifulSoup / Requests	Fetches articles from 50+ ethnic media sources on a 6-hour scheduler
NLP Classifier	Keyword engine + TF-IDF / Logistic Regression ML	Scores and categorizes every article across 5 threat categories
Alert System	Python / JSON pipeline	Builds multi-channel alert packages; routes HIGH/CRITICAL for human review
AI Counter-Narrative	OpenAI gpt-5-mini via Replit integration	Generates culturally-sensitive response messaging on demand
Web Portal	Flask / HTML / Tailwind CSS / Chart.js	Secure, role-gated dashboard for analysts, community partners, and admins

Data Storage

File / Location	Contents
data/raw/raw_articles_master.json	All scraped articles (unclassified)
data/classified/classified_articles.json	Articles with risk scores, categories, matched indicators
data/dashboard_payload.json	Processed payload served to the analyst dashboard
data/pending_alerts.json	HIGH/CRITICAL alerts queued for human review
data/alerts/	Approved alert packages (JSON, organized by timestamp)
data/ml_model/tfidf_classifier.pkl	Trained TF-IDF + Logistic Regression model
data/audit_logs/audit_YYYY-MM-DD.json	Daily audit logs — all platform events
data/ml_seed_articles.json	86 curated training examples (all 5 threat categories)

3. End-to-End Data Pipeline

Every piece of content CIRP processes travels through a deterministic, auditable pipeline. The pipeline can be triggered automatically on the 6-hour scheduler or manually by an analyst via the dashboard Refresh button.

Step 1 — Ingestion (Scraping)

- The Scraper Scheduler (runs every 6 hours) calls the backend scraping engine.
- Each configured source is fetched via HTTP. Article title, content, URL, publication date, language, and target community are extracted.
- Articles are deduplicated by URL and appended to data/raw/raw_articles_master.json.
- Each article receives a unique article_id (MD5 hash of URL + title).

Step 2 — NLP Classification

- Every unclassified article is passed through the NarrativeClassifier.
- Phase A — Keyword Scoring: The article title and content are scanned against a multi-language keyword dictionary covering all 5 threat categories. Each match adds to a weighted keyword_score.
- Phase B — ML Classification: The TF-IDF vectorizer transforms the text and the Logistic Regression model predicts a category and confidence score.
- Phase C — Score Blending (3 modes): (A) If ML agrees with keyword category at ≥45% confidence → score boost applied. (B) If keywords are silent (<2.0) but ML is confident (>60%) → ML drives the score and category. (C) If keywords are weak (<5.0) but ML is very confident (>72%) in a different category → ML category overrides.
- Coordinated Inauthentic Behavior (CIB) detection flags clusters of articles with suspiciously similar content from different sources.
- NER (Named Entity Recognition) extracts countries, organizations, and persons mentioned.
- Final output: risk_score (0–10), risk_level (MINIMAL/LOW/MEDIUM/HIGH/CRITICAL), threat_category, matched_phrases, alert_agencies.

Step 3 — Payload Generation

- Classified articles are sorted by risk score (descending).
- Template-based counter-narrative stubs are attached to each threat article.
- Community-level aggregations, trend data, and language bias stats are computed.
- All data is written to data/dashboard_payload.json for the portal to serve.

Step 4 — Alert Routing

- The alert system loads all threats from the dashboard payload.
- For each non-MINIMAL threat, a multi-channel alert package is assembled.
- CRITICAL and HIGH threats are automatically written to data/pending_alerts.json with status 'pending' — awaiting human review before deployment.

- MEDIUM and LOW alerts are dispatched immediately to the appropriate community channels.
- Each alert package carries a TLP classification: AMBER (government), GREEN (community media/network), WHITE (social platforms).

Step 5 — Dashboard Presentation

- The Flask server reads the payload and serves it to authenticated analysts.
- The dashboard displays threat cards, risk distribution charts, geographic bubble maps, trend graphs, and the pending review badge.
- Analysts can open any threat to view full classification details, matched indicators, alert agencies, and the counter-narrative panel.

4. Threat Classification Engine

4.1 — Five Threat Categories

Category ID	Display Name	What It Covers	Example Signals
foreign_adversary_warfare	Foreign Adversary Warfare	State-sponsored disinformation, election interference, [illegible]	"stolen election", "foreign agent", "deep state", "secret [illegible]
criminal_exploitation	Criminal Exploitation	Human trafficking lures, fraud, money laundering, exploitative job offers targeting [illegible]	"guaranteed job abroad", "send processing fee", [illegible]
social_cohesion	Social Cohesion Threats	Narratives designed to divide ethnic communities, provoke violence, distrust in institutions	"race war", "invasion", "replacement", "us vs them", [illegible]
health_disinfo	Health Disinformation	Vaccine misinformation, fake cures, anti-science health claims, COVID/epidemic-related false [illegible]	"vaccine kills", "5G causes illness", "miracle cure", [illegible]
unknown	Unclassified	Articles that do not match any category with sufficient confidence — flagged for analyst review	Broad or ambiguous content with no clear threat signal

4.2 — Risk Scoring & Level Thresholds

Each article receives a continuous risk_score from 0.0 to 10.0. The score is composed of a normalized keyword component (using an exponential decay function) plus an ML confidence boost. The table below shows how scores map to action levels:

Risk Level	Score Range	Threat Type	Deployment Channels	Human Review
	8.0 – 10.0	Active foreign interference / mass harm potential	Government (TLP:AMBER) · Community Media · Social Monitoring · Community Network	Required
HIGH	6.0 – 7.9	Coordinated disinformation / criminal exploitation	Government · Community Media · Social Monitoring	Required (auto-queued)
	4.0 – 5.9	Emerging harmful narrative / unverified claim	Community Media · Community Network	Optional
LOW	2.0 – 3.9	Low-signal content, worth monitoring	Community Network (prebunking advisory)	Not required
	0.0 – 1.9	No actionable threat signal detected	None — logged for trend analysis only	Not required

4.3 — Coordinated Inauthentic Behavior (CIB) Detection

CIRP implements a basic CIB detector that flags clusters of articles where ≥3 articles from different sources share highly similar content (Jaccard similarity > 0.4) within a 48-hour window. CIB-flagged articles receive a score multiplier of 1.5× and are automatically escalated to HIGH priority if they would otherwise score MEDIUM.

5. Alert System & Multi-Channel Deployment

CIRP generates structured alert packages in four deployment channels, each with its own Traffic Light Protocol (TLP) classification and intended audience:

Channel	TLP	Recipients	Format	Triggers At
Government	AMBER	FBI Cyber Division, DHS/CISA, DoD Cyber Command, HSI, DOJ, HHS, [illegible] Education Security	Structured JSON inter-agency alert [illegible] indicators and recomme	HIGH + CRITICAL
Community Media	GREEN	Ethnic Media Services, NAHJ, NABJ, SAJA, New America Media	Press Advisory with fact-check resources and editorial guidance	MEDIUM + above
Social Monitoring	WHITE	Platform Trust & Safety Teams, GIFCT, First Draft Partners	Platform policy report with matched content and IOC signals	HIGH + CRITICAL
Community Network	GREEN	Local liaisons, faith leaders, cultural center coordinators	Prebunking brief with talking points and community reporting channel	LOW + above

5.1 — Human Review Checkpoint

HIGH and CRITICAL alerts are not dispatched immediately. They are written to a pending review queue (data/pending_alerts.json) with status 'pending'. A red pulsing badge in the analyst dashboard header shows how many alerts are waiting. An admin must log into the admin panel, review the full alert package, and either approve (triggering deployment to data/alerts/) or dismiss the alert. All review actions are recorded in the audit log.

5.2 — Alert Deduplication

Before queuing, the system checks whether an article_id already exists in pending_alerts.json. If a duplicate is detected, the new entry is silently dropped. This prevents the same article from being reviewed multiple times across scrape cycles.

6. AI Counter-Narrative Generation

CIRP integrates OpenAI's gpt-5-mini model to generate culturally tailored counter-narratives for each detected threat. The system is designed to assist — not replace — human analysts and community communicators.

How It Works

1. An analyst opens a threat article in the dashboard and clicks 'Deploy Counter-Narrative'.
2. The button enters a loading state. The article title, content, threat category, targeted community, risk level, and source are sent to POST /api/counter-narrative.
3. A structured prompt is constructed that instructs the AI to produce JSON containing: suggested_response, fact_check_sources (authoritative URLs), recommended_tone, and cultural_note.
4. The OpenAI model returns a JSON response. The counter-narrative panel in the modal is updated live with the AI-generated content, tagged 'AI-Generated' in the UI.
5. If the OpenAI call fails for any reason (network, rate limit, etc.), a template-based fallback response is returned automatically — the analyst always sees a result.
6. The generation event is logged in the audit log with the article title and whether AI or template was used.

Prompt Design Principles

- Community-aware: The prompt includes the target_community field so the model tailors tone and references appropriately (e.g. referencing the Ministère de la Santé for Haitian health disinformation vs. CDC for English-language content).
- Non-amplifying: The system prompt explicitly instructs the model never to repeat harmful frames or restate the original narrative.
- Actionable: Responses are written for field use — community liaisons, ethnic media editors, or government briefers — not for internal analysis.
- JSON-constrained: response_format is set to json_object to ensure structured, parseable output every time.

7. Analyst Dashboard — Features & Capabilities

Login & Authentication

- JWT-style session stored in sessionStorage. Credentials validated server-side against hashed user records.
- Failed logins are rate-limited. All login events (success and failure) are written to the audit log.
- Session automatically expires; logout clears all stored tokens.

Stats Overview Bar

- Total Threats, Critical/High count, Average Risk Score, Active Alerts — all computed live from the classified payload.
- A red pulsing 'N Pending Review' badge appears in the header when HIGH/CRITICAL alerts are awaiting admin approval.

Threat Cards Grid

- Each classified article above MINIMAL is rendered as a card showing: title, source, language, risk level badge (color-coded), risk score bar, threat category, targeted community, and detection date.
- Cards are sorted by risk score descending. Analysts can filter by risk level, category, language, and community.
- Clicking a card opens a full threat detail modal.

Threat Detail Modal

- Displays full classification details: risk score, threat category, matched keyword phrases, alert agencies, CIB flag, ML classification and confidence.
- Shows the counter-narrative panel with pre-loaded template response.
- Three action buttons: Download Intel Report (JSON), Deploy Counter-Narrative (live AI), Share Intel.
- Deploying counter-narrative calls the live AI API and replaces the panel content in real time.

Analytics Charts

- Risk Distribution (doughnut chart): breakdown of articles by risk level.
- Threat Category Breakdown (horizontal bar): article volume by category.
- Top Active Sources (horizontal bar): which ethnic media outlets have the most flagged content.

Geographic Spread & Velocity

- Bubble map showing 35 diaspora community hotspot cities. Bubble size reflects article volume; color reflects highest risk level.
- Velocity Tracking panel shows recent scrape cycles and article counts per source.

Pattern Intelligence — Trend Panel

- Time-series chart showing threat volume and risk level distribution over rolling 30-day windows.
- Escalation flags when a category shows >2 standard deviations above baseline.
- Cross-cycle trend detection to identify sustained campaigns vs. one-off spikes.

Access Request Form

- Non-authenticated users can request portal access. Request is logged and routed to the admin panel for review.

8. Admin Panel & Governance

The Admin Panel (/admin) is a separate, password-protected interface accessible only via the X-Admin-Token header or the admin login form. It provides platform governance and oversight across five functional tabs:

Tab	Key Functions
Users	View all registered users, edit tier/role, deactivate accounts, approve access requests
Access Requests	Review and approve/deny portal access requests from community partners
Alert Reviews	Inspect HIGH/CRITICAL alert packages pending human review; approve (deploy) or dismiss with notes
Audit Log	Browse daily audit logs filterable by event type and date; event counts by category
System Status	ML model health (training date, sample count, per-category distribution), language bias monitor, retrain button, scraper

Audit Event Types

Event Code	Triggered When
login_success / login_failure	User authenticates or fails authentication
access_request	New portal access request submitted
article_analyzed	An article is processed by the classification pipeline
alert_queued	A HIGH/CRITICAL alert is written to the pending review queue
alert_approved	An admin approves a pending alert for deployment
alert_dismissed	An admin dismisses a pending alert with review notes
counter_narrative_generated	An analyst generates an AI counter-narrative
ml_retrained	An admin triggers a full ML model retrain

9. User Roles & Access Control

Role Tier	Label	Access Level	Typical User
government	Government Analyst	Full dashboard read access; can generate counter-narratives; can share intel	Federal/state agency analyst
government_reviewer	Government Reviewer	Above + can access alert review queue; can provide review annotations	Supervisory analyst, TL
government_admin	Government Administrator	Full platform access including admin panel and user management	Platform administrator
community	Community Partner	Dashboard read; threat cards (MEDIUM and below); cannot see government alert package	NGO, ethnic media editor
community_reviewer	Community Reviewer	Above + can submit review annotations and flag articles for re-analysis	Senior community partner

User accounts are provisioned by a government_admin tier user. All provisioning and role-change events are logged in the daily audit file. Community tier users see a government-clearance badge on their session to indicate access limitations.

10. Security, Privacy & Compliance

Control	Description
Authentication	Session-based authentication with hashed credentials. All login attempts logged. Sessions expire on browser clo
Authorization	Role-based access control (RBAC) with 5 tiers. Admin panel protected by separate API token (X-Admin-Token h
Data Handling	All scraped content is public-facing ethnic media. No personally identifiable information (PII) is collected or stored
TLP Classification	All alert packages carry Traffic Light Protocol markings: AMBER (government only), GREEN (community partner
Audit Trail	Every platform action (logins, article analysis, alert review, counter-narrative generation, ML retrain) is timestamp
Human Review Requirement	HIGH and CRITICAL alerts cannot be deployed without explicit admin approval — preventing automated dissem
AI Transparency	Every AI-generated counter-narrative is tagged 'AI-Generated' in the UI and logged with the model identifier. Ten
Retention	Raw articles and audit logs are retained indefinitely. Pending alerts preserve full review history including reviewe

11. Glossary

CIB — Coordinated Inauthentic Behavior — clusters of similar content from different sources suggesting orchestrated amplification.

TLP — Traffic Light Protocol — a standard for classifying the sensitivity and sharing scope of intelligence: RED (recipient only), AMBER (organization), GREEN (community), WHITE (public).

TF-IDF — Term Frequency–Inverse Document Frequency — a statistical text feature extraction method used to train the ML classifier.

LR — Logistic Regression — the supervised machine learning algorithm used to classify articles into threat categories.

NER — Named Entity Recognition — NLP technique to extract entities (people, organizations, locations) from article text.

Keyword Score — A weighted sum of matched threat-indicator phrases, normalized on a 0–7 scale using an exponential decay function.

Risk Score — The final 0–10 threat score combining keyword score and ML confidence boost.

Pending Alert — A HIGH or CRITICAL alert that has been generated but not yet approved by a human reviewer.

Counter-Narrative — A messaging strategy designed to proactively rebut a harmful narrative using trusted, authoritative sources without amplifying the original content.

Diaspora Media — News and media outlets that serve immigrant and diaspora communities, often published in heritage languages.

Mode A / B / C — Three ML blending modes: A (ML corroborates keywords), B (ML detects what keywords miss), C (ML overrides weak keyword category signal).

Dashboard Payload — The pre-computed JSON file (data/dashboard_payload.json) that the portal serves to analysts — updated after each pipeline run.