Functional Requirements Specification

1. **Data Ingestion & Scraping Module**
   1.1. **Social Media Integration (API-Driven):**
   1.1.1. **Twitter, Facebook, Instagram:** Utilize official APIs exclusively to scrape posts, replies, user profiles, hashtags, timestamps, geographical data (if available), and associated media (images, videos). Implement robust mechanisms for API access, rate limiting, and strict compliance with each platform's terms of service, including comprehensive error handling for API failures. **Direct web scraping of social media via headless browsers or similar techniques is explicitly forbidden due to legal, ethical, and stability concerns.**
   1.1.2. **Multi-Lingual Support:** Capable of ingesting and identifying content in all major global languages, supporting Unicode and diverse character sets.
   1.2. **Web Page Scrapping:**
   1.2.1. **URL Management:** Allow administrators to define and manage a list of specific web page URLs for continuous or scheduled scraping, ensuring multi-lingual content extraction.
   1.2.2. **Content Extraction:** Extract main post content, comments, author information, timestamps, and embedded multi-modal media from specified web pages, including support for diverse language encodings.
   1.2.3. **HTML Parsing:** Utilize robust HTML parsing techniques resilient to page structure changes and varying linguistic presentations.
2. **Hate Speech Identification & Trigger Mechanism**
   2.1. **"Reported Post" Definition:** "A 'reported post' is defined as any post on a monitored social media platform to which at least one reply post contains a designated #hate-speech hashtag. While the 'reported post' itself typically does not contain this hashtag, it might occasionally include a designated #hate-speech hashtag, potentially as a deceptive mechanism to evade detection or create confusion. The system's primary trigger mechanism remains the hashtag in the reply."
   2.2. **Sole Trigger Mechanism: The presence of a configurable designated #hate-speech hashtag in a *reply post* is the *sole* trigger for initiating deeper analysis of the *original "reported post"* (the post being replied to).**
   2.3. **No Proactive Unreported Content Analysis:** If a social media post potentially contains hate speech but no user has reported it by replying with the designated #hate-speech hashtag, the system is explicitly *not* to analyze or process it. The rationale is to prevent giving undue exposure to unreported hate speech and to manage the formidable task of analyzing all social media content by focusing only on reported instances.
   2.4. **No Analysis of Reporting Replies:** Reply posts containing the #hate-speech hashtag are solely considered as a reporting mechanism and are *not* subject to hate speech analysis themselves. The system focuses its analysis exclusively on the "reported post" that the hashtagged reply targets.
   2.5. **Hashtag Configuration Management:**
   2.5.1. **Trustee Control:** The specific set of #hate-speech hashtags (which can be multi-lingual, e.g., #hate-speech, #discoursehaineux, #odioonline) for each language is managed and determined by the designated international "Anti-Hate Speech Trustees."
   2.5.2. **On-Demand Updates:** Updates to these hashtags are expected to occur on-demand as determined by the trustees.
   2.6. **Multi-threaded Monitoring Process:** A dedicated multi-threaded process will continuously search monitored social media channels for the presence of these configured #hate-speech hashtags in reply posts. Once identified, the process retrieves the corresponding "reported post" (the parent of the reply) for further ingestion and analysis.

3. **Anonymous User Submission & Multi-Modal Content Ingestion Module (CORE FEATURE)**
   3.1. **Anonymous Submission Interface:**
   3.1.1. **Secure Portal:** Provide a highly secure, anonymous web portal where users can submit evidence without revealing their identity (implementing robust anonymity protocols and technical safeguards).
   3.1.2. **Multi-Lingual Interface:** The submission interface itself must be available in multiple languages.
   3.2. **Multi-Modal Content Upload:**
   3.2.1. **URL Submission:** Allow users to submit URLs to web pages, social media posts, or online media files suspected of containing hate speech. **A user-submitted URL for a social media post will be treated as a web page for scraping purposes, bypassing the hashtag-based trigger for that specific instance.**
   3.2.2. **Audio File Upload:** Enable the secure upload of audio files (e.g., MP3, WAV) up to a specified size/duration.
   3.2.3. **Video File Upload:** Enable the secure upload of video files (e.g., MP4, MOV) up to a specified size/duration.
   3.2.4. **Image File Upload:** Enable the secure upload of image files (e.g., JPEG, PNG) up to a specified size.
   3.3. **Metadata Capture:**
   3.3.1. For each submission, allow users to optionally provide contextual metadata (e.g., approximate date/time of event, brief description of hate speech, relevant keywords, perceived location of event). All metadata fields should be optional to maintain anonymity and ease of submission.
   3.4. **Content Sanitization & Virus Scanning:**
   3.4.1. Implement immediate and robust virus and malware scanning on all uploaded files.
   3.4.2. Sanitize uploaded content where technically feasible to remove hidden metadata that could compromise anonymity (e.g., EXIF data from images).
   3.5. **Temporary Storage:** Securely store uploaded content in an encrypted temporary storage area prior to analysis.
4. **Hate Speech Analysis Module (Enhanced for Multi-Modal & Multi-Lingual)**
   4.1. **Initial Assessment & Flagging:**
   4.1.1. Upon ingestion (either from a social media "reported post" or an anonymous user submission), the content will be automatically analyzed by NLP/ML models to raise it as a "potential hate-speech" instance.
   4.1.2. **Automated Multi-Lingual Detection:** Utilize advanced NLP, machine learning, and deep learning models (e.g., CNNs for images, RNNs for audio/video transcripts) to initially assess content and determine if it unequivocally *could* contain hate speech across all supported languages and modalities.
   4.1.3. **Text Analysis:** For social media, web pages, and transcribed audio/video.
   4.1.4. **Image Analysis:** Detect symbols, gestures, text within images, and visual cues associated with hate speech.
   4.1.5. **Audio Analysis:** Transcribe spoken language and analyze tone/intonation for markers of aggression or hate.
   4.1.6. **Video Analysis:** Transcribe audio, analyze visual content (gestures, symbols, context), and detect textual overlays.
   4.1.7. **Confidence Scoring:** Assign an initial confidence score to each detection of *potential* hate speech.
   4.2. **International Anti-Hate Speech Trustees Review & Voting:**
   4.2.1. All "potential hate-speech" instances are recorded and presented to the "International Anti-Hate Speech Trustees" via the platform's UI.

4.2.2. **Up/Down Voting:** Trustees are empowered to vote instances "up" or "down" based on their expert judgment.
4.2.3. **Dynamic Threshold Activation:** If a "potential hate-speech" instance reaches a pre-defined *dynamic threshold* scoring (e.g., adjusted based on number of "up" votes from trustees, the number of times the post has been reported and confirmed by analysis as a hate-speech post, or the high average severity score set by trustees), it is then formally identified, recorded, and classified by the system as confirmed "hate speech."
4.3. **Thorough Analysis for Confirmed Hate Speech:**
4.3.1. Only *after* an instance has been formally identified as "hate speech" by the trustees' voting process, will the system proceed with a comprehensive, "thorough analysis" to extract detailed contextual information. This includes all items listed below:
4.3.2. **Multi-Lingual Entity Extraction:**
4.3.2.1. **Target Identification:** Automatically identify the specific target(s) of the hate speech (individuals, groups, ethnicities, religions, genders, nationalities) in all supported languages.
4.3.2.2. **Subject/Topic:** Determine the overarching subject or topic of the hate speech.
4.3.2.3. **Accusation/Claim:** Extract specific accusations, claims, or derogatory statements made.
4.3.3. **Associated Hashtags/Keywords:** Identify all relevant hashtags or keywords mentioned in textual content.
4.3.4. **Temporal Data:** Accurately extract the date and time of the reported content's creation or the reported event.
4.3.5. **Geographic Location:** Extract geographic location information if available (e.g., from post metadata, user profile, EXIF data if anonymized, or inferred from content).
4.3.6. **"Hate-Manufacturer" / Source Identity:** Identify the creator of the reported post/content (username, profile link, unique ID, or "Anonymous User Submission" for submitted content).
4.3.7. **Jurisdiction Mapping (Global):**
4.3.7.1. **Manufacturer/Source Jurisdiction:** Attempt to identify the likely legal jurisdiction of the "hate-manufacturer" or the origin of the hate speech.
4.3.7.2. **Involved Jurisdictions:** Identify other jurisdictions potentially implicated by the content, target, or reported event.
4.3.8. **Named Entity Recognition (NER - Multi-Lingual):**
4.3.8.1. **Mentioned People:** Extract a list of named individuals.
4.3.8.2. **Mentioned Characters:** Extract fictional or public figures referred to.
4.3.8.3. **Mentioned Organizations:** Extract named entities representing organizations.
4.3.8.4. **Mentioned Events:** Extract named events.
4.3.9. **Victim Identification:** Identify specific entities or groups explicitly targeted as victims within the hate speech content.

5. **Database Management Module**
5.1. **Schema Design (Global & Multi-Modal):** Design a robust, flexible SQL database schema (tables, fields, relationships) capable of securely storing all extracted information from diverse sources and modalities, including:
5.1.1. Reported Post/Submission Details (content, URL, platform, timestamp, manufacturer ID/anonymity flag, language, original modality, transcription text, **initial potential hate-speech flag**).
5.1.2. Hate Speech Analysis Results (**trustee voting results, confirmed hate speech flag**, confidence, categories, review status, **results of thorough analysis**).
5.1.3. Extracted Entities (targets, subjects, topics, accusations, hashtags, named entities)
5.1.4. Jurisdictional Data (multiple per record)
5.1.5. User and Authentication Data (for platform users, including Trustees)
5.1.6. Aggregation Records
5.1.7. Report Records

5.1.8. Web Page Records (for moderated pages)
5.1.9. Voting Records (for public scoring)
5.1.10. Audit Logs for all data access and modifications.
5.2. **Data Persistence:** Store all processed data securely and efficiently in the SQL database, ensuring data integrity and availability across geographical regions.
5.3. **Index Optimization:** Ensure appropriate indexing for fast retrieval, aggregation queries, and multi-lingual searches.
5.4. **Transaction Management:** Implement ACID-compliant transactions for all database operations.
5.5. **Multi-Lingual Search Indexing:** Optimize the database for efficient multi-lingual search and retrieval of content and entities.

6. **Aggregation & Reporting Module**
6.1. **Web User Interface (UI) - General & Multi-Lingual:**
6.1.1. **Intuitive & Responsive Design:** Provide a user-friendly, modern, and efficient UI for all platform functions, available in all supported languages.
6.1.2. **Role-Based Access Control (RBAC):** Implement distinct UI views and functionalities based on user roles (e.g., Administrator, Analyst, International Anti-Hate Speech Trustee, Public Viewer).
6.1.3. **Localization:** Support for various date/time formats, number formats, and cultural conventions.
6.2. **Trustee-Specific Interface:** A dedicated section within the UI for Trustees to review "potential hate-speech" instances and cast their votes.
6.3. **Search & Filtering (Advanced & Multi-Lingual):**
6.3.1. Allow users to search and filter database records based on any extracted field, including language, modality of submission, hate-manufacturer, victim, topic, date range, jurisdiction, specific hashtags, confidence score, and **official "hate speech" classification status**.
6.3.2. Support complex boolean queries and fuzzy matching.
6.4. **Aggregation Workbench (Enhanced):**
6.4.1. **Interactive Aggregation:** Provide sophisticated tools within the UI to intuitively group and aggregate records from all sources based on:
6.4.1.1. Same event (temporal proximity, shared entities, geographical correlation)
6.4.1.2. Same hate-victim (individual or group)
6.4.1.3. Same hate-manufacturer/source (including anonymous aggregate analysis)
6.4.1.4. Same topic/theme (cross-lingual topic modeling)
6.4.1.5. User-selectable set of tags/criteria
6.4.1.6. Jurisdictional overlap
6.4.2. **Multi-Modal Summarization:** Offer summaries that can incorporate excerpts from text, images, audio transcripts, or video snippets.
6.4.3. **Visualization:** Offer powerful, interactive graphical representations and statistical summaries of aggregated data, capable of handling large, diverse datasets.
6.5. **Aggregation Output Options:**
6.5.1. **Save as Data Record:** Persist aggregated findings as a new, distinct record in the database for further analysis.
6.5.2. **Generate Report:** Create formal, customizable, and exportable reports (e.g., PDF, CSV, DOCX) from aggregated data, suitable for legal submission or international bodies. Reports must be configurable with various templates and languages.
6.5.3. **Display as Web Page:** Publish aggregated findings as a dynamic, shareable, multi-lingual web page within the platform.

7. **Public Exposure & Scoring Module (New)**
7.1. **Public Web Page Display:** Once an aggregated report is published as a web page, it becomes publicly viewable through the platform's interface.

7.2. **Public Hate-Speech Scale Scoring:** Allow *any* user (who must be authenticated) to score the displayed web page on a "hate-speech scale" (e.g., raise or lower its score points by one point), indicating their perception of its importance and severity.
7.3. **Dynamic Listing & Reporting:**
7.3.1. Web pages will be listed and referenced in descending order of their current score on the public hate-speech scale.
7.3.2. **United Nations Security Council Consideration:** The platform should highlight or specifically present the "Top 10" highest-scoring hate-speech web pages for consideration by bodies such as the United Nations Security Council, with the understanding that this is a developing feature and its full implications will evolve.

8. **Data Replication & Recovery Module (Globally Distributed & Secure)**
8.1. **Real-time Replication (Geographic Redundancy):**
8.1.1. Implement a separate, continuous process to copy database records and track all changes (inserts, updates, deletes) from the primary database to one or more geographically dispersed replica databases. Specify if replication should be synchronous for critical data and asynchronous for less critical data.
8.1.2. Ensure data consistency and integrity across all replicas.
8.2. **Transaction Tracking & Immutable Log:**
8.2.1. Log all database transactions (DDL and DML) in an immutable, cryptographically secured, auditable log.
8.3. **Point-in-Time Recovery:**
8.3.1. Enable the ability to roll back the source database to any specific transaction point, ensuring data integrity.
8.4. **Special Case Backup for Rollback & Legal Archiving:**
8.4.1. **Pre-Synchronization Backup:** Before synchronizing the source and replica databases after a rollback, automatically create a special, legally compliant, and cryptographically signed backup archiving the distinct state of *both* the source and replica databases at that exact moment. This ensures distinct historical states are preserved for legal and auditing purposes, potentially across multiple storage locations.