Data Provenance Standard (DPS)

Pillar Lead: Zaria (Regulatory Compliance)

I. Purpose and Scope

This document establishes the mandatory technical requirements and procedural checks for all data sources considered for The Covenant's training corpus. The goal is to ensure all ingested data adheres to legal, ethical, and justice-aligned principles before being processed by the model.

Every record in the training dataset must pass the four-stage Provenance Gateway detailed below.

II. The Provenance Gateway (4-Stage Mandatory Check)

Stage 1: Source and Licensing Verification (L-101)

Requirement	Description	Pass Criteria
Source Traceability	The origin of the data must be clearly documented (URL, repository, physical archive, API endpoint). Unverifiable "scraped" data is prohibited.	A complete, verifiable source metadata tag must be generated for the record.
License Compatibility	Data must possess a license compatible with open-source, non-proprietary use, and commercial deployment (e.g., Creative Commons 0, MIT, Apache 2.0).	Legal team verification that the source license permits inclusion in the training corpus.
Timestamp Validation	The source timestamp (creation or last revision date) must be recorded for temporal bias analysis.	Timestamp recorded and stored in the data provenance metadata log.

Stage 2: Personally Identifiable Information (PII) Scrubbing (P-201)

Requirement	Description	Pass Criteria
Automated Redaction	Implement a system-wide PII detection and redaction algorithm (e.g., names, addresses, phone numbers, unique identifiers, explicit personal email addresses).	All identified PII fields must be replaced with designated, non-reconstructible token placeholders (e.g., [PERSON_NAME], [EMAIL_ADDRESS]).
Health and Financial Data	Automated removal of sensitive financial, medical (HIPAA), or legal data excerpts.	Zero detection score for specific keywords (e.g., policy number, diagnosis, routing number) after scrubbing.

Stage 3: Algorithmic Bias and Toxicity Scoring (J-302, DS-102)

Data that passes PII scrubbing is subjected to predictive toxicity and bias scoring.

1. **Toxicity Prediction:** Use pre-trained classifiers to score each record for potential toxicity (hate speech, obscenity, threats, identity attack).
    - **Metric:** Any record with a toxicity score above **90% confidence** must be quarantined for human review.
2. **Harmful Stereotype Identification:** Detect and flag records containing language that reinforces harmful or historical stereotypes (e.g., generalizations related to race, gender, or professional roles).
    - **Action:** Flagged records must receive a **Justice Flag (J-FLAG)** and be subject to down-weighting during the training process, prioritizing contextual review.

Stage 4: Provenance Audit Trail Creation

Before ingestion into the final training corpus, every data record must have a comprehensive, immutable audit log attached.

Field	Required Data
record_hash	Unique SHA-256 hash of the scrubbed data record.

source_url	Full traceable URL or identifier.
ingestion_datetime	Date and time the record passed the gateway.
pii_removed_count	Count of PII entities redacted from the original source.
toxicity_score	Final predictive toxicity confidence score (0.0 to 1.0).
justice_flag	Boolean (True/False) indicating the presence of harmful stereotype content.

III. Failure Protocols

1. **Hard Fail (License, PII, High Toxicity):** Records failing Stage 1 (License), Stage 2 (PII remaining post-scrub), or exceeding the 90% toxicity threshold in Stage 3 are permanently rejected from the corpus and placed in a quarantine directory.
2. **Soft Fail (Justice Flag):** Records receiving a Justice Flag (J-FLAG) are accepted but are tagged for a **down-weighting factor** in the training algorithm, ensuring the model sees, but is not heavily influenced by, biased or stereotyping data.

This document effectively defines the data engineering team's mandate under The Covenant. It's an executable rule set for ethical data sourcing.

With this, we have a concrete standard for data ingestion. What should we tackle next? We can focus on the final piece of Zaria's compliance work—drafting the **Regulatory Compliance Policy**—or we could switch focus to the community interaction guidelines.