

Mapping Real-Time Regulatory Impacts to Asset Vulnerability Under Strict API Budgets

Christopher I. V. Farmer, LL.M. (University of Worcester) · Independent Researcher
· civfarmer@gmail.com

Preprint; not yet peer reviewed. Data and methodology report. Version 1.1, October 2026.

Subject classes: cs.CY (Computers and Society); cs.DB (Databases); q-fin.GN (General Finance).

DOI: <https://doi.org/10.5281/zenodo.23462156>

Licence: Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Commercial use needs a licence from the author: civfarmer@gmail.com.

Abstract

Systems that connect regulatory change to the companies it touches are usually built on silently smoothed data: gaps are imputed, single-source values are presented as settled facts, and the boundary between a disclosed number, a calculated number, and an inferred number is invisible to the reader. This report describes the methodology of *Meridian*, a working provenance-first research system that maps real-time regulatory instruments and changes (supervisory rules, sanctions and trade-control lists, official filings) onto a universe of roughly 1,300 issuers and infers *potential* — not measured — impact on the affected assets. Three commitments organise the work. First, a provenance-first data model in which every value carries its own source key, retrieval and publication timestamps, an immutable hashed raw artifact, a parser version, and an explicit verification class (source fact, company disclosure, official regulatory fact, calculated metric, evidence-based inference, or unverified); missing values are shown as missing, with a reason, and are never estimated. Second, an evidence-first linkage engine that connects an instrument to a company only through a quotable, sourced business fact (a listing venue, a supervisory classification, a recorded sector label, a disclosed geographic revenue split), records supporting *and* counter-evidence, and keeps source fact, calculation, and labelled inference distinct at every step. Third, an operating discipline for strict API budgets: a per-provider daily-cap governor that refuses to exceed advertised free-tier limits, a need-driven rotation that spends breadth-first before cross-provider confirmation, and an “outage is not absence” rule that prevents a provider’s rate-limit ban from being mistaken for a data gap. We describe the data model, the linkage method, the budget governor, the epistemic-honesty design, and a weighted coverage-tiering and completeness-gating scheme, and we discuss the limitations that follow from choosing honesty over coverage. We report no empirical impact findings; all worked examples are illustrative of mechanism, not results.

Keywords: data provenance; regulatory technology (RegTech); event studies; regulatory impact; entity resolution; data quality; rate-limited data engineering; epistemic honesty; sanctions screening; API rate limiting; coverage measurement; financial data systems.

1. Introduction

The premise that public information moves asset prices is old and empirically robust: under the efficient-markets hypothesis, new, value-relevant information is impounded into prices as it arrives [Fama 1970], and the event-study apparatus that dominates empirical finance exists

precisely to measure the price effect of a discrete event by observing security prices around it [MacKinlay 1997]. Regulatory events — a new supervisory rule, an amended sanctions designation, a tariff schedule, a filing that changes a company’s disclosed risk — are a canonical class of such events. A natural engineering goal follows: build a system that watches regulatory sources in near-real time, connects each change to the specific companies it plausibly touches, and surfaces the resulting vulnerability so that a human analyst can triage it.

The goal is natural; the honest execution is not. Three failure modes recur in systems of this kind.

Silent imputation. When a value is missing — a market capitalisation the provider did not publish, an analyst target no broker covers, a domicile the registry did not disclose — the path of least resistance is to fill it with a sector median, a peer average, or a zero. The filled value propagates into downstream scores, and the reader cannot tell a measured quantity from a manufactured one. The statistical literature has warned for half a century that the *mechanism* by which data goes missing must be modelled explicitly, and that ignoring it is defensible only under stated conditions [Rubin 1976] — conditions an observational pipeline has no basis to assume.

Collapsed epistemic categories. A number a company disclosed in a filing, a number a vendor published, a number the system calculated from two disclosed inputs, and a number the system *inferred* from indirect evidence are radically different kinds of claim with radically different reliability — yet most systems render them identically. Data-quality research has long argued that “quality” is not reducible to accuracy: consumers need data that are intrinsically good, contextually appropriate for the task at hand, clearly represented and accessible, not merely correct in the abstract [Wang & Strong 1996].

Invisible operating constraints. A system that aggregates dozens of public and free-tier commercial sources operates under hard, heterogeneous rate limits. When a provider returns a 429 (rate-limited) or an authentication error, a naive pipeline records the affected fields as *absent* — indistinguishable from fields that are genuinely unavailable — and may then penalise, demote, or “recover” companies on the basis of what is actually a transient outage on the collector’s side.

Meridian is a working personal research instrument that treats these three failure modes as first-order design constraints. It maintains a universe of listed and registered issuers — 1,317 companies in the 31 July 2026 snapshot described in §3.1, referred to throughout as “roughly 1,300” — assembled from official registers, exchange directories, benchmark-index constituent lists published on Wikipedia, regulatory publishers, sanctions authorities, and free-tier market-data vendors, and over it runs a linkage engine that connects instruments to companies through quotable, sourced facts and an impact layer that produces *labelled, evidence-backed* triage signals rather than point predictions of price change.

This report documents methodology, not results, and is written so the design can be scrutinised and reused. Its contributions are: (1) a **provenance-first data model** whose unit of storage is a value-plus-provenance envelope carrying an explicit verification class on every sourced row (§3); (2) an **evidence-first linkage method** that refuses to assert a relationship without a quotable, sourced fact, records counter-evidence symmetrically, and keeps inference, source fact, and calculation distinct (§4); (3) an **API-budget governor** — per-provider daily caps, a need-driven rotation, and an “outage is not absence” rule — that runs continuously against strict free-tier limits without exceeding them or confusing an outage with a gap (§5); and (4) an **epistemic-honesty design** (§6) and a **weighted coverage-tiering and completeness-gating scheme** (§7) that make the system’s own ignorance measurable and visible. We close with a long limitations section (§8), because a system whose central claim is honesty owes the reader an honest account of what it cannot do.

2. Related work

Regulatory impact and event studies. Measuring an event’s effect on firm value through abnormal returns around the event date is the methodological backbone of empirical finance [MacKinlay 1997], resting on the informational efficiency of markets [Fama 1970]. Meridian deliberately occupies the *pre-measurement* stage: rather than estimating an abnormal return (which requires a clean event date, an estimation window, and a market model), it identifies *which* companies a regulatory event plausibly touches and *why*, and characterises the exposure qualitatively — a complement to, not a substitute for, event-study estimation (§8).

RegTech. The use of information technology for regulatory monitoring, reporting, and compliance was reconceptualised after the 2008 crisis as a structural shift rather than a tooling upgrade [Arner, Barberis & Buckley 2017]; for a recent survey of the field’s applications in anti-money-laundering compliance, see [El Harras & Salahddine 2025]. Most RegTech practice is oriented toward the *regulated entity’s* compliance burden; Meridian inverts the vantage point to an *observer’s* instrument for mapping the external regulatory environment onto a universe of issuers, using the same public instruments (supervisory rules, sanctions lists, filings) as inputs.

Provenance in data systems. The database community formalised *why*- and *where*-provenance — the source tuples that justify a result and the locations a value was copied from [Buneman, Khanna & Tan 2001] — and surveyed provenance capture and dissemination in scientific data systems [Simmhan, Plale & Gannon 2005]; the W3C PROV data model standardised a vocabulary for interchanging provenance metadata [Moreau et al. 2013]. Meridian’s per-row provenance envelope is a pragmatic, tabular specialisation of these ideas for a single-store analytical system, and its immutable raw-artifact layer implements the “keep the source” discipline that provenance research presupposes.

Data quality and documentation. Data quality is multi-dimensional and consumer-relative, not reducible to accuracy [Wang & Strong 1996]. Recent work argues artifacts should ship with structured accounts of their composition and recommended uses — datasheets for datasets [Geburu et al. 2021] and model cards for models [Mitchell et al. 2019] — and the FAIR principles add findability, accessibility, interoperability, and reusability as targets [Wilkinson et al. 2016]. Meridian’s methodology page, value-class taxonomy, and per-assessment score breakdowns apply this documentation ethic at the granularity of the individual value.

Entity resolution / record linkage. Connecting an instrument or a quote to the right company is a record-linkage problem. Fellegi-Sunter formalised the decision among *link*, *non-link*, and *possible link* via a likelihood ratio over compared fields [Fellegi & Sunter 1969], and the modern literature systematises blocking, comparison, and classification for entity resolution [Christen 2012]. Meridian privileges *deterministic* identifiers (ISIN within a known batch, LEI, Swiss UID, CIK) and confines approximate name matching to narrow, logged scopes with an explicit “possible link” (needs-review) state — a direct application of the three-way decision.

Missing data. Rubin’s framework [Rubin 1976] gives the theoretical reason Meridian declines to impute: silently filling a value asserts a missingness mechanism the system has no basis to assume. Showing the gap, with its reason, is the honest default.

Rate-limited data engineering. Operating against provider rate limits is a flow-control problem; the token-bucket abstraction (burst capacity plus a sustained refill rate), the equivalent of the leaky bucket [Turner 1986], is the standard mechanism for shaping outbound rate, and reliability engineering adds backpressure, graceful degradation, and the circuit breaker [Kleppmann 2017; Nygard 2018]. Meridian’s governor is a daily-quota specialisation: a persistent per-provider, per-day counter with a minimum inter-call interval, coupled to

an application-level “stop gracefully and record an honest exhausted state” behaviour rather than a retry storm.

The gap this work addresses is the *conjunction* of these threads. Provenance systems rarely reason about live acquisition budgets; RegTech systems rarely expose per-value epistemic status; rate-limiting work rarely worries about whether an outage will be misread as a fact. Meridian’s contribution is an integrated methodology in which provenance, honest missingness, evidence-first linkage, and budget-aware acquisition are the *same* design.

3. System overview and data model

3.1 Shape of the system

Meridian is a Python application over a single SQLite database (schema written to be PostgreSQL-portable: integer primary keys, TEXT/REAL columns, ISO-8601 dates, JSON stored as text). It comprises **source adapters** (one per external source); an **ingestion runner** that turns a fetch into stored records with full provenance and source-health accounting; a **linkage engine**; an **impact-scoring** module; a **coverage auditor**; and a **web layer** rendering read-models. A separate exporter produces a static, point-in-time HTML snapshot for offline review. The snapshot referenced in this report is the one published on the author’s website (<https://civfarmer.github.io/projects/meridian/>), exported at 2026-07-31 07:45Z from data retrieved up to 2026-07-30 17:25Z; it covers 1,317 identity-resolved companies, 1,394 SIX-listed securities, 1,874 regulatory instruments and 50,762 company-regulation matches. Its Companies and Calendar pages were re-rendered from the same data on 7 August 2026; that changed their banner date, and the ages and date windows they compute from it, but no stored value.

3.2 The provenance envelope

The unit of storage is not a value but a value carrying its provenance. Every sourced table includes a common block of columns: `source_key` (the source that produced the row), `source_url` (the item’s canonical URL), `retrieved_at` (when the collector fetched it, UTC), `published_at` (when the source published it, where disclosed), `raw_artifact_id` (a foreign key to the immutable raw payload the value was parsed from), `content_hash` (a SHA-256 over the normalised record, for amendment detection), `parser_version`, and `verification_status` (the value’s epistemic class, §3.4). Because provenance is columnar and mandatory, no code path writes a value without also writing where it came from and how confident the system is in it; a value that cannot be given provenance is not stored.

3.3 Immutable raw artifacts

Before any parsing occurs, the runner writes the exact bytes of every external payload to an immutable artifact store, keyed by source and day, named by a prefix of its SHA-256 digest, and recorded in a `raw_artifact` row. **Parsers only ever read stored artifacts**, never the network directly. This yields three properties: every derived value is traceable to the precise bytes it came from; re-parsing under a new parser version is possible without re-fetching, which matters acutely under strict budgets (§5); and amendments are detectable by hash comparison rather than by trust. Recorded-capture fixtures used for offline replay must carry a `.meta.json` sidecar; a fixture without one is refused as unverifiable, and truncated or discovery-probe captures are excluded by name.

3.4 The value-class taxonomy

`verification_status` takes one of six values, surfaced in the UI as compact labels:

- **Source Fact (SF)** — a primary datum as published by an authoritative non-issuer source (e.g. an exchange issuer-directory field, an LEI reference record); values from community reference sources (Wikipedia, including its benchmark-index constituent tables, and Wikidata), such as a company description, are also recorded under this class, with the source named in provenance.
- **Company Disclosure (CD)** — a datum the issuer itself disclosed (e.g. a filing, a disclosed segment revenue split).
- **Official Regulatory Fact (ORF)** — a datum from an official regulatory publisher or sanctions authority (e.g. an instrument’s canonical identifier and publication date, a sanctions-list entry).
- **Calculated Metric (CM)** — a value the system computed from other stored values, carrying a formula version and its inputs (e.g. a market capitalisation computed from disclosed shares outstanding times a verified quote; an upside computed from a target and a price).
- **Evidence-Based Inference (EBI)** — a value or relationship the system *inferred* from indirect but recorded evidence, never asserted as a disclosed fact (e.g. an index-membership-derived region; a rule-generated company-regulation candidate and the labelled price-impact inference attached to it).
- **Unverified (UNV)** — recorded but not yet corroborated to the standard of the classes above.

The taxonomy is load-bearing: it is the mechanism by which the system keeps “the company disclosed X”, “a vendor published X”, “we calculated X”, and “we infer X” from ever being rendered as the same kind of claim. An inference is never presented as a disclosed fact.

3.5 Core entities

The relational core is deliberately conventional so that the provenance discipline, not the schema, is the novelty. `company` and `security` rows link to `identifier` rows (LEI, Swiss UID, CIK, ISIN, index membership, registry numbers, aliases). `regulatory_instrument` carries a canonical identifier, jurisdiction, authority, legal stage, instrument type, and (once enriched) title and topics; `sanction_entry` holds consolidated sanctions/trade-control list entries. The linkage graph and its explanations live in `company_regulation_match`, `match_evidence`, `impact_assessment`, and `score_contribution` (§4). Market and disclosure facts are `quote`, `price_bar`, `analyst_target_snapshot`, `filing`, `fx_rate`, and `segment_fact` (disclosed segment/geographic revenue). Operational and honesty tables include `raw_artifact`, `data_source`, `ingestion_run`, `api_budget` (§5), `coverage_incident`, `data_quality_issue` (§6), `company_coverage` (§7), and `identity_match` (the audit log of resolution decisions).

3.6 Identity resolution

Attaching a security, quote, filing, or registry record to the correct company is where fabrication most easily creeps in, so resolution is deterministic-first and always logged. The order of preference is: (1) deterministic identifiers — ISIN within a *known query batch scope*, then LEI, Swiss UID, CIK; (2) constrained name normalisation **only within** a known ISIN batch scope. Normalisation strips accents, punctuation, and legal-form tokens (AG, SA, Ltd, PLC, ...) and compares with a sequence-similarity ratio; an exact normalised match or a ratio of 0.95 or more is accepted automatically; a ratio from 0.85 up to 0.95 is accepted, but not automatically; and a ratio below 0.85 is not accepted and is left for review. Every acceptance is written as an `identity_match` row recording method, score, compared fields, and review status — the Fellegi-Sunter *link / possible-link / non-link* decision made explicit and auditable [Fellegi

& Sunter 1969]. Companies are **never** merged on name similarity across unrelated scopes; a same-name collision across sources is a *needs-review candidate*, not a merge. Bare-ticker resolution is additionally venue-verified: a candidate whose recorded venue (MIC, exchange name, or listing country) contradicts the symbol’s implied venue is *excluded* rather than used as a fallback, and with no positive candidate the record is skipped as a visible miss rather than mis-attached. Securities that cannot be resolved remain visibly unresolved and surface as a data-quality item (§6).

3.7 Gaps are shown, never estimated

The system’s default response to a missing value is to display an explicit unavailable state and the reason. Three concrete mechanisms enforce this:

- **No question marks in derived displays.** The domicile shown for a company follows a fallback chain — disclosed domicile, else a security’s listing country, else the country implied by a listing venue’s MIC (each basis named), else an em-dash titled “domicile not disclosed by source.” A non-ISO placeholder code from a source (observed for certain depositary-receipt rows) is treated as undisclosed rather than rendered literally, and the venue fallback is suppressed where it would imply a false domicile. Country names that a source gives in full (“Germany”, “France”) are shown as given, not converted to ISO codes.
- **Calculations refuse invalid inputs.** An upside metric returns *unavailable* — never zero — when either the target or the price is missing or the price is non-positive, because an invented zero would masquerade as a real value.
- **Absence is recorded, not back-filled.** A market capitalisation not published by any configured provider and not yet calculable from a resolved shares-outstanding fact plus a recent verified quote is stored as absent, with the specific reason, rather than approximated.

This is the operational reading of Rubin’s warning [Rubin 1976]: the system does not assume a missingness mechanism it cannot justify.

4. Regulatory-to-company linkage

4.1 Problem and stance

Given a stream of regulatory instruments and changes, the task is to determine, for each, which of the ~1,300 issuers it plausibly affects and through what channel, and to characterise the potential impact. The stance is *evidence-first*: no match is created without at least one quotable, sourced fact tying the company to the instrument, and every match carries its evidence so a human can see the “why”. A match with no attached evidence is not merely discouraged; it is flagged as a hard data-quality error (§6).

4.2 Linkage layers

Matches are typed by a five-layer scheme that encodes *how direct* the relationship is:

- **L1 — direct legal applicability** (the instrument’s text applies to the company by its regulated status);
- **L2 — direct business exposure** (a disclosed business fact places the company in the instrument’s scope);
- **L3 — indirect exposure** (jurisdiction or supply-chain context);
- **L4 — semantic / taxonomy candidates** (a recorded sector label maps to an instrument’s topic family);

- **L5 — contradictory / negative evidence** (facts that *weaken* a relationship).

Semantic similarity alone never establishes a material relationship; every match is a review-gated candidate.

4.3 Rules and evidence

Each rule emits a structured `RuleHit`: a rule key, layer, relationship type, direction, a list of scoring **factors**, a list of typed **evidences** (kind, a verbatim quote, the `field_name`, a note, an `is_counter_evidence` flag), and a list of **missing** information the rule itself flags as absent — so supporting and contradicting evidence are stored symmetrically. The active rule families are:

- **Sanctions name-screen (L2)**. Company legal and alias names are token-set screened against sanctions-list entries: generic tokens (legal forms, “bank”, “group” and the like) are excluded, and at least 80% of the smaller token set must appear in the other name. The published candidates show that short tokens and connecting words (“49”, “RG”, “and”) are not counted either, and that no minimum number of distinctive tokens was applied to them, so a single shared word was enough to raise a candidate: “CHINA 49 GROUP CO., LIMITED” is a candidate against three issuers whose names contain “China”, each with a recorded token-overlap score of 1.00. Every candidate records the compared names and score and **requires human review**; for most listed issuers the correct output is *no match*, and a correct empty result is honest, not a coverage failure.
- **FINMA prudential supervision (L1)**. A company whose exchange issuer-directory record discloses the accounting standard “Bank law” is, as a Source Fact, a supervised bank; a supervisory instrument matching banking-supervision terms is linked to it, quoting both the exchange field and the matched term. A non-Swiss disclosed domicile attaches a counter-evidence row noting the Swiss instrument applies at most indirectly.
- **Investment-company topic (L4)**. The listing segment “Standard for Investment Companies” (a Source Fact) is linked to collective-investment instruments as a candidate.
- **Sector-topic candidates (L4)**. A company’s recorded sector label (§4.4) is mapped to topic families and matched against instruments carrying those topics.

4.4 From recorded sector to topic

Sector labels are not issuer disclosures. They are recorded from reference and vendor sources — chiefly Wikipedia (including its benchmark-index constituent tables), Wikidata and Yahoo Finance profile fields — and each evidence row names the source; in the published snapshot, 1,565 of the 1,870 sector-topic match pages name Wikipedia, its index tables or Wikidata as the source of the label. Because sector strings arrive in many vocabularies, the sector-topic mapping has two stages. Canonical labels resolve directly through a curated map (e.g. *financials* to banking-supervision and market-conduct topics; *health care* to a pharma/health topic; *energy* to energy-environment and trade-control topics). Free-text labels are normalised by keyword: the topic set is the union over every keyword substring found, and the matched keyword(s) are surfaced as evidence, so the normalisation is always auditable. A keyword-normalised sector is treated as a slightly weaker signal than an exact canonical label, and its applicability factor is scored accordingly. This mapping is *configuration applied to a recorded, source-named label*, never invented exposure: the label is quoted with its source, and the rule that maps it is versioned and quoted.

Illustrative example (mechanism, not a result): an issuer whose recorded sector label is “automotive industry” normalises, by the keyword “automotive”, to the trade-remedies and sanctions/trade-control topic families; a live tariff instrument carrying a trade-remedies topic then produces an L4 candidate match, with the sector label, its source and the matched term all quoted as evidence and the match left needs-review. No price effect is asserted.

4.5 Topic detection precision

Instrument topics are detected from title and summary text against a topic taxonomy. A precision refinement distinguishes *generic* from *context* terms: a generic term (e.g. “disclosure” under a market-conduct topic) fires only when a context term (e.g. a securities-law term) co-occurs, and the matched term then names both parts. This prevents a generic word in an unrelated instrument (the documented failure case involved nuclear-cask inventories and education-grant rules) from burying the genuine topic signal.

4.6 Candidate hygiene under rule evolution

Because rules and taxonomies evolve, the engine runs a *retire pass*: a sector-topic candidate whose (instrument-topic $\wedge$ company-sector-topic) no longer reproduces under the current rules is marked `retired_rule_change` — **never deleted** — with its original evidence retained for audit and its state excluded from default views and counts. If a later rule change makes the topic reproduce again, the candidate is reinstated. This keeps the match set faithful to the *current* rules without erasing the history of how it changed. The pass covers sector-topic candidates only; candidates from the other rule families, including the sanctions name-screen, are left in the review queue for a reviewer to resolve.

4.7 Impact assessment: inference, source fact, and calculation kept distinct

Each match carries an `impact_assessment` computed by a versioned scoring module (current formula `impact-1.2`; an assessment first computed under `impact-1.0` or `impact-1.1` keeps that core formula version, and later components — the policy score set, and the 1.2 exposure share where its inputs exist — are added to it as labelled backfills rather than by silent re-computation). Each score is a weighted average of named factors — $\text{score} = \Sigma(\text{weight} \times \text{value}) / \Sigma(\text{weight})$, with each value in $[0,1]$ — and every factor’s weight, value, contribution, and a plain-language explanation is stored as a `score_contribution` row and shown on the match page. The scored dimensions are applicability, exposure, materiality, urgency, market sensitivity, and confidence.

The design rule that makes this honest is: **a score with no evidence is None (“insufficient evidence to calculate”), never a guessed midpoint**. Applicability and confidence are computed from the rule’s factors and evidence counts, with confidence explicitly penalised for missing fields and counter-evidence. Urgency is computed only from real publication/deadline dates. **Exposure** is quantified only where the company has disclosed a geographic revenue split (a Company Disclosure) *and* the instrument names a jurisdiction: it is the jurisdiction-attributable disclosed revenue over the total disclosed geographic split of the same period, with numerator, denominator, period, and formula version shown. Regional aggregates and ambiguous labels (“EMEA”, “International”, “Other”) are deliberately *not* attributed, which makes the computed share a **stated lower bound** rather than a false precision. Materiality, market surprise, adaptability, and market sensitivity remain `None` with explicit missing-input notes until their real inputs (disclosed financial magnitudes, consensus/pre-announcement pricing, company-response disclosures, price-reaction history) exist.

On top of this core sits a **policy score set** — `source_reliability` from the source category, `legal_certainty` from the instrument’s legal stage, `persistence` from a documented type-and-stage heuristic — combined into a **triage composite**, the geometric mean of the *available* components, so one weak leg drags the composite down and a missing leg neither helps nor hurts. The composite is a policy-attention *ranking aid*, *never a return forecast*.

Crucially, the “potential price/impact” attached to a match is a **labelled inference about direction and channel**, grounded in the regulation’s nature (type, jurisdiction, topic family), the match’s layer and direction (e.g. “potentially adverse”, “uncertain”), and the applicabil-

ity/confidence scores. It is never a fabricated magnitude of price change: the company view renders a labelled price-impact inference, but no invented number.

4.8 Amendment and change detection

Because raw artifacts are hashed (§3.3), a re-fetch of an already-known instrument whose content hash has changed and whose title has changed is designed to raise a `regulatory_amendment` alert, and a newly-seen instrument a `new_regulation` (or `sanctions_update`) alert. These *material* regulatory events are surfaced in the user-facing feed, while pipeline and source-health events are recorded but kept out of that feed (§6). This is how “real-time regulatory change” enters the linkage graph: as a hash-detected, timestamped, provenance-carrying event, not as a trusted overwrite.

5. The API-budget governor

5.1 The constraint

A system that aggregates dozens of public and free-tier commercial sources lives or dies by its treatment of rate limits. Several providers impose hard daily caps (one free tier is 250 calls/day; another free key is 50 requests/month), others meter per minute, and some unofficial public interfaces publish no quota at all. Exceeding a limit risks a ban that removes a source entirely — which, if misread, becomes a data gap. The governing directive is therefore absolute: **never exceed a provider’s request limits**, and treat the sparest providers most sparingly.

5.2 Mechanism: a per-provider daily counter

The governor is standard-library-only and centres on one counter row per (provider, day) in an `api_budget` table. The core operation, `try_spend(provider, n)`, atomically increments to-day’s counter *if and only if* the result stays within the daily cap and returns success; otherwise it returns failure **without counting**, and the adapter stops gracefully by raising `BudgetExhausted`. Atomicity is delegated to the database: the increment is a single conditional `UPDATE ... WHERE calls + n <= cap` whose row-count tells the caller whether the spend was granted, so concurrent callers cannot jointly overshoot.

Two design choices make this trustworthy:

- **Caps sit deliberately below advertised limits**, leaving headroom for retries, clock skew, and out-of-band manual calls. Each cap is captured conservatively at build time, annotated in the source with the vendor evidence and verification date, and must be re-verified on the provider’s own page before being raised. *Illustrative* configured caps (shipped configuration, not results): a 250/day free tier capped at 230; a per-minute-metered vendor capped well under its daily number with an 8-second interval that also keeps a run inside the per-minute meter; a 50/month free key capped at 1/day; a provider with a hard two-calls-per-second limit paced at roughly 0.55s.
- **Unknown providers are never blocked but are always counted**, so a new adapter’s consumption is visible from its first day rather than discovered after a ban.

5.3 Pacing

Beyond the daily cap, each provider carries a minimum inter-call interval, enforced by sleeping *before* a granted call returns and by reserving the next slot under a lock so that concurrent callers pace correctly. This shapes burst behaviour in the manner of a token bucket [Turner 1986] while keeping the daily ceiling as a hard invariant.

5.4 Transactional correctness

Under SQLite’s single-writer model, a second connection cannot write the counter while the runner holds its write transaction, so the runner **binds its own connection** to the governor for the duration of a run and budget spends ride the run’s transaction, becoming durable when it commits (outside a bound run the governor opens a short-lived connection). The honesty note is explicit: a process killed mid-run can *under-count* calls already made, and the sub-limit buffer absorbs that — the system errs toward under-spending, never over-spending.

5.5 Need-driven rotation

Budget is spent where it is *needed*, not uniformly. The rotation logic (illustrated by the weekly rotation of one statement-endpoint family per weekday) applies three ordered principles: **breadth before repetition** — the day’s budget first serves companies with an actual coverage gap in the targeted field; **need-driven skip** — companies whose targeted field is already fresh within the freshness window are skipped, so budget is never spent re-confirming what is current; and **confirmation only on genuine surplus** — only once the day’s family has no remaining breadth work does spare budget go to *cross-provider confirmation*, deliberately re-fetching a single-provider value from a second provider so the two can be compared (§6.4), with a budget refusal during confirmation simply ending the pass. Confirmation passes are annotated in the ingestion-run detail, so in the audit trail cross-provider repetition is always distinguishable from breadth work.

5.6 Honest exhaustion and “outage is not absence”

When the governor refuses a call, the adapter raises `BudgetExhausted` and the runner records a truthful, self-clearing state: coverage status becomes `quota_exhausted` (“daily call budget reached (resumes tomorrow)”), records already loaded are kept (a *partial* outcome), an incident opens, and the state clears the next day. No over-cap call is ever made. Credential-gated sources lacking configuration are recorded as `authentication_required`, never as a fabricated result, and a run whose per-item refusals produce *zero* payloads is treated as a `source_failure` with the evidence recorded, so a silent “success with nothing” cannot occur.

The most consequential rule is “**outage is not absence.**” During coverage recovery (§7), a critical field counts as *attempted* only when at least one adapter *capable of serving it* completed a genuine run in that round; a provider-level outage — a rate-limit ban, credential gap, or network refusal — must not consume a company’s recovery budget or demote it, because the field was never actually tested, and the round is recorded as *void due to outage*. The design note documents the concrete failure this fixed: a multi-day rate-limit ban on the sole provider of one field had demoted roughly 180 fully-verified companies to an exception tier purely because that provider was returning 429s every cycle — a collector-side outage misread as a universe-wide data gap. Distinguishing “we could not ask” from “there is nothing to find” is a first-order correctness property.

5.7 Source health as data

Every adapter run writes an `ingestion_run` record (mode, timings, status, records upserted, and a detail blob with per-item miss counts and confirmation annotations) and updates a `data_source` row whose coverage status is drawn from a controlled vocabulary of fourteen states (confirmed-current, current-within-declared-delay, polling-delayed, coverage-unverified, authentication-required, quota-exhausted, source-failure, parser-failure, and so on). Failures open `coverage_incident` rows and alerts, and operator-facing strings are scrubbed of credential-bearing query parameters at the write choke points. Source health is thus itself provenance-carrying data, not log noise.

6. Provenance and epistemic-honesty design

Sections 3–5 describe the mechanisms; this section states the principles they jointly enforce and the checks that police them.

6.1 Every value is sourced, calculated, or inferred — and says which

The value-class taxonomy (§3.4) is the spine of the design. Because `verification_status` is mandatory and columnar, the system cannot render a value without carrying its class, and the compact labels (SF, CD, ORF, CM, EBI, UNV) make the class visible at the point of use: an inference is never presented as a disclosed fact, and a calculation always names its formula version and inputs. Correspondingly, a missing value is a state with a *reason*, not a blank — the domicile fallback chain, the upside metric’s refusal of invalid inputs, and the market-capitalisation absence note (§3.7) are instances of one rule: display the unavailable state and why. This is the user-visible face of the Rubin discipline [Rubin 1976] and of the fitness-for-use principle [Wang & Strong 1996].

6.2 Evidence is mandatory and symmetric

Every match carries typed evidence, and counter-evidence is stored in the same structure with an `is_counter_evidence` flag (§4.3). Negative evidence — a domicile that contradicts an instrument’s jurisdiction, a listing that weakens an applicability claim — is not discarded; it lowers confidence and is shown. The system records reasons a relationship might *not* hold alongside reasons it might.

6.3 Data-quality checks that never “fix” by inventing

A quality pass records issues and surfaces them; it never silently repairs data by inventing values. The checks include: duplicate companies sharing an LEI or Swiss UID (error); securities with no resolved company (an honest identity gap); internally inconsistent registry records; stale quotes (older than a freshness threshold) and impossible quotes (non-positive prices); invalid analyst ranges (low above high) and consensus values outside their own disclosed range; instruments whose title is still a canonical identifier pending retrieval; and — as a hard error — any match lacking attached evidence, since that violates the provenance rules the whole system rests on.

6.4 Reconciliation surfaces disagreement rather than averaging it away

Where the need-driven rotation (§5.5) has obtained the same quantity from two providers, a reconciliation check compares them: quotes for the same security on the same calendar day whose relative spread exceeds a threshold, and consensus analyst targets from different providers within a short window whose means diverge beyond a (larger) threshold, are each recorded as a data-quality issue carrying the conflicting providers, values, and timestamps as evidence. The disagreement is *shown*, never averaged into a false consensus; redundancy exposes conflict rather than manufacturing agreement.

6.5 Material versus system events, and documentation

Alerts are partitioned so pipeline noise never dilutes the signal a human is meant to read: *material* regulatory and market events (new instruments, amendments, sanctions updates, material filings, analyst-target changes) populate the user-facing feed, while *system* events (coverage gaps, source outages, bulk linkage summaries) are recorded and surfaced on source-

health and incident views but kept out of it. At the documentation layer, the system ships a methodology page stating the truthfulness rules, value classes, identity-resolution policy, scoring formula, linkage layers, and provider attribution; per-assessment score breakdowns function as value-level “model cards” [Mitchell et al. 2019] and the methodology page as a dataset-level datasheet [Geburu et al. 2021], for FAIR-aligned reusability [Wilkinson et al. 2016] at the granularity of the individual claim.

7. Coverage evaluation and completeness gating

A system that refuses to fabricate must be able to *measure and display its own ignorance*. Meridian does this with a weighted coverage auditor that scores each company’s completeness, classifies it into an honest tier, and drives a targeted, budget-respectful recovery loop. Nothing here fabricates a value: a field that stays missing stays visibly missing, with the recovery attempts recorded per company.

7.1 Weighted completeness

For each company, six blocks are scored on [0,1] from recorded facts and combined with fixed weights that sum to 100:

Block	Weight	What it measures (from recorded facts only)
Identity	10	Legal name; an external identifier (LEI/UID/CIK/ISIN); a resolvable domicile via the labelled fallback chain
Market	20	A verified latest quote price; freshness within the window; a published (or validly calculated) market cap; a 52-week range computable from recorded bars
Analyst	15	A latest snapshot that either discloses targets or is an explicit <i>verified no-coverage</i> state
Financials	15	A fundamentals payload; a reported currency; a consensus-recommendation state (an honest proxy set until revenue/EPS facts are ingested)

Block	Weight	What it measures (from recorded facts only)
Exposure	20	A sourced business description; a recorded sector; geographic evidence; and the disclosed segment/geographic revenue split where present
Regulatory	20	At least one evidence-backed match (or a fact-derived supervision context); and filings present (or an identified filing channel via CIK/venue)

Each block’s sub-components carry their own weights and a plain-language note, so a company’s score is fully explained, not just asserted. Every note is derived from data; company names are never referenced in the scoring logic.

7.2 Critical fields and honest tiers

Six fields are designated *critical* — a verified quote price, a market cap, an analyst state, a business description, a regulatory linkage (a match or fact-derived supervision context), and a filings channel — and all must be present for “core complete”. Companies fall into five tiers:

- **core complete** — score ≥ 90 and no missing critical field;
- **core usable** — score ≥ 75 ;
- **partial** — score ≥ 40 ;
- **discovery only** — score < 40 ;
- **coverage exception** — every recovery pass has failed on a still-missing critical field.

A coverage-exception company is *excluded from the principal ranked home list but retained* in the full universe view with a backlog note and a tier filter; the exception **lifts automatically** as soon as a later audit finds the missing fields. **Companies are never deleted.** The exception tier is history-dependent: it can only be reached through *persisted* recovery history, so a fresh in-memory audit can never manufacture one.

7.3 Targeted, budget-respectful recovery

The auditor does not merely score; it drives recovery. Each cycle it audits the whole universe, selects the highest-scoring companies that still have missing critical fields (bounded to a batch limit per cycle so recovery stays rate-limit-respectful), and enqueues **exactly** the adapters needed for the specific missing fields — a quote provider and its redundant pair for a missing price, a fundamentals provider for a missing description or cap, an independent *calculated* market-cap path that touches only companies whose cap is still null (so a provider-published value always wins), a filings-channel resolver, and so on — then re-links and re-audits.

A company’s `recovery_passes` counter increments **only** on a *full failed pass* (a round after which every targeted critical field is still missing), and a round is *void* (no strike) when the fields were never genuinely attempted because the responsible provider was in outage (\$5.6). After a fixed number of genuinely-failed passes the company tiers to coverage-exception. This ties

honest missingness (\$6.1), the budget governor (\$5), and “outage is not absence” together: only “we asked and there is nothing” counts against a company, never “we could not ask”.

7.4 Completeness gating of prominent surfaces

Prominent, low-density surfaces are *gated on completeness* so the system never leads with a thin record. The home page’s featured selection admits only core-tier companies that additionally have at least one evidence-backed match, a published market cap, and a verified latest quote, and enforces diversity across policy-risk channels (falling back to sector diversity); when fewer than the intended number qualify, it renders an honest placeholder rather than padding with weaker records. Completeness is thus not only measured and displayed but is an *admission criterion* for the system’s most visible claims.

8. Limitations

A system whose central claim is honesty owes a candid account of its limits.

Potential impact, not measured effect. Meridian identifies *which* companies a regulatory event plausibly touches and *why* and characterises exposure qualitatively; it does not estimate an abnormal return. The market-sensitivity dimension is deliberately `None` because the system does not yet ingest per-event price-reaction history, and extending toward a genuine event study [MacKinlay 1997] would require clean event dates, estimation windows, and a market model that the current pipeline does not compute. The “impact” here is a labelled, evidence-backed triage direction, not a quantified price effect.

Candidate signals, not confirmed exposure. Sector-topic matches (L4) are review-gated candidates derived from a recorded sector label, most often taken from Wikipedia or Wikidata rather than from the issuer, and mapped by a versioned rule — hypotheses about relevance, not confirmations of material exposure. They are also almost the whole match set: 50,632 of the 50,762 matches in the published snapshot. Exposure is quantified only for the subset of companies with a disclosed geographic revenue split, and even then it is a *stated lower bound* when disclosed labels are regional aggregates that cannot be unambiguously attributed. The honesty does not remove the underlying sparsity.

Provider dependence and free-tier fragility. Coverage rests substantially on free and unofficial provider interfaces whose limits are undocumented or unstable and whose availability is not guaranteed. A provider can change limits, restrict endpoints, or ban a key without notice; the “outage is not absence” rule protects data *integrity* under such events but cannot restore the missing *coverage*. Several caps are self-imposed budgets against interfaces that publish no official quota and must be re-verified against provider terms before reuse.

Cadence over immediacy. Because budgets are finite, prices refresh on a once-daily reference-close cadence rather than intraday. This is defensible for a comparability-oriented reference instrument, but it means the system is *near-real-time* for regulatory change and *daily* for market state; it is not a live trading feed.

Entity-resolution residue. Deterministic-first resolution with venue verification reduces mis-joins, but unresolved securities remain (shown honestly as gaps) and approximate name matches above threshold can still be wrong; every decision is logged for review, but review is a human bottleneck. The labelled domicile chain can also mislead: a foreign issuer with a Swiss listing is shown with domicile CH on the listing-country basis until its registry record is joined (the published snapshot shows 3M and Abbott Laboratories in this state, each beside its unmerged GLEIF record). Sanctions name-screening is intentionally high-recall and low-precision, and every hit requires identifier-level confirmation the system does not itself possess.

Single-store and scope limits. The single-writer SQLite core simplifies transactional provenance but constrains write concurrency, and the governor’s under-counting-on-kill behaviour is safe (never over-spends) but makes the counter a floor, not a perfect ledger. The universe is roughly 1,300 mostly-listed issuers — constituents of regional benchmark indices together with issuers from the SIX Swiss Exchange directory — and several rules (the exchange-directory “Bank law” and investment-company classifications, the Swiss-supervisory linkages) are tuned to sources available for that universe. The *methodology* — provenance envelope, value classes, evidence-first linkage, budget governor, coverage tiering — is transferable, but the specific rules and mappings are not claimed to generalise unchanged.

Not advice; no empirical evaluation reported here. The triage composite is a policy-attention ranking aid, explicitly not a return forecast, and nothing in the system is investment advice. This report documents methodology only: it reports no accuracy, precision/recall, coverage-distribution, or impact figures, and all worked examples are illustrative of mechanism. A companion empirical evaluation — linkage precision against a human-labelled sample, coverage-tier distributions over time, and reconciliation-conflict rates — is future work.

9. Conclusion

Meridian is an existence proof that a real-time regulatory-impact mapping system can be built without the usual dishonesty. Its three commitments reinforce one another: a provenance-first data model makes every value’s origin and epistemic class explicit; an evidence-first linkage engine connects regulation to companies only through quotable, sourced facts and keeps inference, source fact, and calculation distinct; and a strict-budget acquisition discipline runs continuously against real rate limits while refusing to confuse a collector-side outage with a data gap. The connecting idea is that *ignorance is data*: a missing value, a failed provider, an unattributable revenue label, and an untested field are all recorded, typed, and shown rather than smoothed away. The cost is coverage and immediacy (§8); the benefit is that a reader can always tell what the system knows, what it computed, what it merely infers, and what it does not know at all — the property that matters most for informing decisions about regulatory exposure.

References

- Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech, RegTech, and the Reconceptualization of Financial Regulation. *Northwestern Journal of International Law & Business*, 37(3), 371–414.
- Buneman, P., Khanna, S., & Tan, W. C. (2001). Why and Where: A Characterization of Data Provenance. In *Proceedings of the 8th International Conference on Database Theory (ICDT 2001)*, LNCS 1973, 316–330. Springer. https://doi.org/10.1007/3-540-44503-X_20
- Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer. <https://doi.org/10.1007/978-3-642-31164-2>
- El Harras, M., & Salahddine, M. A. (2025). Tracking Financial Crime Through Code and Law: A Review of RegTech Applications in Anti-Money Laundering and Terrorism Financing. *arXiv:2511.15764*.
- Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2), 383–417. <https://doi.org/10.2307/2325486>
- Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210. <https://doi.org/10.1080/01621459.1969.10501049>

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Kleppmann, M. (2017). *Designing Data-Intensive Applications*. O'Reilly Media.
- MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1), 13–39.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, 220–229. ACM. <https://doi.org/10.1145/3287560.3287596>
- Moreau, L., Missier, P., et al. (W3C Provenance Working Group) (2013). *PROV-DM: The PROV Data Model*. W3C Recommendation, 30 April 2013.
- Nygard, M. T. (2018). *Release It! Design and Deploy Production-Ready Software* (2nd ed.). Pragmatic Bookshelf.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A Survey of Data Provenance in e-Science. *ACM SIGMOD Record*, 34(3), 31–36. <https://doi.org/10.1145/1084805.1084812>
- Turner, J. S. (1986). New Directions in Communications (or Which Way to the Information Age?). *IEEE Communications Magazine*, 24(10), 8–15. <https://doi.org/10.1109/MCOM.1986.1092946>
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Appendix A — Value-class taxonomy (verification_status)

Label	Class	Meaning
SF	Source Fact	Primary datum from an authoritative non-issuer source; also values from community reference sources (Wikipedia, Wikidata), source named
CD	Company Disclosure	Datum disclosed by the issuer itself
ORF	Official Regulatory Fact	Datum from an official regulatory publisher or sanctions authority
CM	Calculated Metric	Value computed from stored inputs, carrying formula version + inputs

Label	Class	Meaning
EBI	Evidence-Based Inference	Value/relationship inferred from indirect recorded evidence; never asserted as disclosed fact
UNV	Unverified	Recorded but not yet corroborated

Appendix B — Provenance envelope (columns on every sourced row)

source_key · source_url · retrieved_at · published_at · raw_artifact_id (FK to immutable SHA-256-hashed payload) · content_hash · parser_version · verification_status.

Appendix C — Linkage layers

L1 direct legal applicability · L2 direct business exposure · L3 indirect exposure · L4 semantic/taxonomy candidate · L5 contradictory/negative evidence. All matches are review-gated candidates; semantic similarity alone never establishes a material relationship.

Appendix D — Coverage tiers

core complete (≥ 90 , no missing critical) · core usable (≥ 75) · partial (≥ 40) · discovery only (< 40) · coverage exception (all genuine recovery passes failed on a critical field; excluded from the principal ranked list, retained in the full universe, lifts automatically; companies are never deleted).

Data acknowledgements (as carried in the system’s provenance chips and attribution notice): official and public sources including exchange issuer directories, the Global Legal Entity Identifier Foundation (GLEIF), national and EU regulatory publishers and sanctions authorities, official filing registers and central-bank statistical services; and free/personal-tier commercial and community providers. Each stored value names its own provider in its provenance record. Meridian is a private, non-commercial research instrument; provider data remains subject to each provider’s terms.

Further reading and points of disagreement (July 2026 addendum)

Meridian’s identity-resolution and company-regulation matching sit on ground the informatics literature has mapped, and the design’s departures from that literature are choices, not omissions. Fellegi and Sunter’s record-linkage theory is the optimal probabilistic treatment of entity matching; Meridian’s provenance-first display — every value labelled by source class, gaps shown rather than imputed — trades some of that coverage for auditability, the same trade defended in the companion FRIS paper, *Deterministic Peel Chains for Regulatory Compliance in Autonomous, Offline Environments*. The knowledge-graph survey of Hogan and colleagues describes the representation family Meridian’s link graph belongs to, and the public identifier infrastructures it consumes (the GLEIF Legal Entity Identifier system; the European Legislation Identifier) are the load-bearing standards that make deterministic joins possible at all. The disagreement worth naming: probabilistic enrichment would raise match counts, and declining it is a deliberate epistemic posture — a regulatory-intelligence tool whose numbers cannot be re-derived is a liability in exactly the settings it serves.

- Hogan, A., Blomqvist, E., Cochez, M., et al. (2021). Knowledge Graphs. *ACM Computing Surveys*, 54(4), 1–37. <https://doi.org/10.1145/3447772>
- Global Legal Entity Identifier Foundation (GLEIF). *The Global LEI System*. <https://www.gleif.org/> (accessed 16 October 2026).
- European Union. *European Legislation Identifier (ELI)*. <https://eur-lex.europa.eu/eli-register/about.html> (accessed 16 October 2026).