Version 1.0 · 2026-05-20 · Public

BA | Intel — Data Methodology

Version 1.0 — May 2026

This document describes how BA | Intel collects, validates, classifies, enriches, retains, and retires on-chain entity labels and sanctions data. It is published as a public reference to support customer due diligence and is suitable as an annex to Data License Agreements (DLA), as evidence under SOC 2 Type II controls, and as part of GDPR Article 30 records of processing activities.

The audience for this document is compliance officers, MLROs, legal counsel, auditors, integrators, and procurement teams evaluating BA | Intel as a data source.

The methodology is the same data backbone used across BA | Oracle (decision-query API), BA | KYT (compliance-grade transaction monitoring), BA | Screen (sanctions and adverse media screening), BA | Investigation (forensic graphs), and BA | API (programmatic access to the suite). All products inherit the rules described here.

1. Purpose, Scope & Intended Use

1.1 What this database is

BA | Intel maintains a multi-chain entity label database — at the time of this version, in excess of 1 billion address-level labels across 52 supported chains — together with a normalized index of crypto-relevant sanctions designations sourced from official government lists. Each label is a structured assertion that a given blockchain address is associated with a specific real-world or behavioral category, accompanied by source attribution, a confidence score, a verification timestamp, and (where applicable) inference tags.

1.2 What it is intended for

BA | Intel data is intended to be used as **evidence** supporting a customer's compliance decision. The customer makes the final determination (block, warn, allow, escalate, file SAR/STR, etc.). BA does not make compliance decisions on behalf of customers.

Intended use cases include:

- Real-time pre-transaction risk screening of crypto addresses (consumer protection and AML).
- Transaction monitoring and ongoing due diligence for regulated VASPs, CASPs, and OTC desks.
- Investigation, fund tracing, and forensic attribution work.
- Sanctions screening obligations under OFAC, EU, UN, UK HMT, and Swiss SECO frameworks.
- Pre-onboarding wallet screening for customer due diligence (CDD).

1.3 What it is not

BA | Intel is not a substitute for the customer's compliance program. It does not constitute legal advice. A label or a sanctions match is a signal that requires interpretation in the context of the customer's policies, risk appetite, jurisdiction, and the specific facts of the transaction.

The database does not contain personal data of identified or identifiable natural persons in the GDPR sense as a default rule. Where on-chain addresses can be linked to natural persons (e.g., via corporate registry crosswalks), such linkage is handled in dedicated downstream products with appropriate lawful-basis assessment and is not exposed via the default Intel surface.

---

2. Sources & Provenance

Every label carries a non-empty `source` field that identifies its origin. Sources fall into seven categories.

2.1 Regulatory and government lists

OFAC SDN (United States), EU FSF Consolidated Financial Sanctions List (European Union), UN Security Council Consolidated List, UK HM Treasury Office of Financial Sanctions Implementation (OFSI) consolidated list, and Swiss SECO SESAM list. These are the only sources that can produce the `SANCTIONED` category (see § 6). Refresh cadence: every four hours via the `sanctions-sync` cron job.

Additionally: enforcement actions published by the U.S. SEC, the UK FCA, and equivalent national regulators are ingested daily as labeled events; these produce `SCAM`, `EXPLOIT`, or `OTHER` categories with appropriate tagging, never `SANCTIONED`.

2.2 Industry blacklists

ScamSniffer (phishing drainers), OpenSanctions (crypto-specific designations crosswalked from official lists and OFAC enforcement actions), Allenhark (Solana rug pull deployers), and community-maintained mixer and darknet-market address sets. Each industry source has a known lag and known false-positive rate, documented in the per-source metadata.

2.3 Exchange attribution

Hot and cold wallets of centralized exchanges, identified via a combination of public disclosure, on-chain pattern matching (deposit address clustering, sweep patterns), and curated lists. Produces the `EXCHANGE` category. Confidence depends on attribution method — direct public confirmation yields the highest confidence; behavioral inference yields lower confidence.

2.4 Token holder extraction

Token contracts with significant adoption (≥ 10,000 holders or strong analytical value) are systematically extracted to label every address that has held the asset. Sources include Etherscan V2 (EVM), Helius

(Solana), TronGrid (TRON), and TON Center. Produces tagged labels (e.g., `["holder", "USDT"]`) under the `OTHER` category — never high-threat categories.

2.5 Behavioral inference

Heuristic patterns derived from on-chain activity: deposit-address patterns, sweep wallets, first-funder analysis, mixer-output detection, cross-protocol behavior. All behavioral inference labels carry the `inferred` tag and an additional method tag identifying the heuristic used. See § 7 for the policy on inferred labels.

2.6 Academic and open datasets

Periodically incorporated datasets from peer-reviewed research, government-published seizure inventories, and verified investigative journalism. Each dataset is reviewed for licensing, accuracy, and methodology before ingestion. Origin and citation are preserved in the `source` field.

2.7 Cluster expansion

Address clustering — particularly UTXO co-spend clustering and SOL program account clustering — expands a single high-confidence attribution to its associated address cluster. All cluster-expansion labels carry the `clustering` tag and are capped at the value-tier and confidence rules in § 5 and § 7.

---

3. Category Taxonomy

Each label is assigned exactly one of the following 17 categories. The taxonomy is a closed enum. New categories may be added in minor versions of this methodology with prior notice (see § 12).

Category	Definition	Typical use
`EXCHANGE`	Centralized exchange hot/cold wallet, deposit address, or operational wallet.	Fund flow attribution
`DEFI`	Decentralized finance protocol smart contract (router, pool, vault).	Protocol exposure analysis
`BRIDGE`	Cross-chain bridge contract or bridge-related address.	Cross-chain risk assessment
`MIXER`	Coin mixer or anonymization service contract.	High-risk AML signal
`GAMBLING`	Gambling protocol or licensed casino wallet.	Jurisdictional risk
`MINING`	Mining pool payout address.	Source-of-funds attribution
`NFT`	NFT marketplace contract or curated NFT collection address.	Collection-specific exposure
`SCAM`	Address attributed to a confirmed scam, rug pull, or fraudulent operation.	High-risk signal
`PHISHING`	Address used in phishing or wallet-drainer campaigns.	High-risk signal
`EXPLOIT`	Address linked to a known exploit, hack, or theft.	Stolen funds tracing
`SANCTIONED`	Address on an official government sanctions list (§ 6).	Mandatory block
`P2P`	Peer-to-peer exchange address.	Higher AML risk
`CUSTODIAL`	Non-exchange custodian (e.g., institutional custody).	Trust assessment
`PAYMENT`	Payment processor or merchant gateway address.	Commercial activity
`WALLET_SERVICE`	Wallet provider or operational address (faucets, smart wallet factories).	Operational attribution
`DAO`	Decentralized autonomous organization treasury or governance address.	Protocol governance
`STABLECOIN_ISSUER`	Stablecoin issuer authorized minting / burning / blacklist address.	Issuer policy exposure
`OTHER`	Anything not fitting the above; almost always accompanied by descriptive tags.	Long-tail attribution

3.1 One label per address per chain

The database enforces `UNIQUE(address, chain)`. The first high-quality label written wins; subsequent inserts use `ON CONFLICT (address, chain) DO NOTHING` and are discarded rather than

overwriting existing labels. This guarantees stability of attribution and prevents low-confidence sources from displacing higher-confidence ones.

---

4. Threat Level Scale & Value Tiers

4.1 Threat levels

Every label carries one of five threat levels:

Threat level	Meaning	Typical categories
`SAFE`	Informational, no risk implication.	`EXCHANGE` (most), `DEFI`, `MINING`
`LOW`	Awareness signal, no action expected.	`NFT`, `WALLET_SERVICE`, `PAYMENT`
`MEDIUM`	Review recommended; may warrant escalation depending on context.	`GAMBLING`, `P2P`, `BRIDGE`, inferred labels
`HIGH`	Strong risk signal; escalation expected.	`SCAM`, `PHISHING`, `EXPLOIT` (direct attribution)
`CRITICAL`	Mandatory blocking signal under applicable law.	`SANCTIONED`

Inferred labels (clustering, 2-hop propagation, cross-chain, first-funder, AI-derived) are capped at threat level `MEDIUM` regardless of the underlying signal strength. See § 7.

4.2 Value tiers

For customer-facing tiering and product gating, labels are further classified into three value tiers:

- **T1 — Compliance Critical.** Labels that drive immediate, legally significant action: `SANCTIONED`, `EXPLOIT` with direct attribution to a known event, `PHISHING` with direct attribution to an active campaign, `MIXER` contracts on official designation lists.
- **T2 — Contextual.** Labels that classify what an address does: token holders, behavioral patterns, exchange attribution, DeFi protocol participation, first-funder relationships.
- **T3 — Coverage.** Long-tail labels providing breadth: NFT collections, smaller DeFi protocols, less-active addresses with single-source attribution.

Value tiers do not override threat levels — a T2 label can be `HIGH` if backed by direct attribution; a T1 label can be downgraded to `MEDIUM` if the underlying source is inferred.

---

5. Confidence Scoring

Every label carries a numeric `confidence` field in the closed interval `[0.0, 1.0]`. The scoring rubric is:

Confidence band	Numeric range	Meaning
Verified	`>= 0.95`	Direct, authoritative source (official list, public disclosure, on-chain proof).
High	`0.80 – 0.94`	Strong evidence from a trusted source; some interpretation involved.
Medium	`0.60 – 0.79`	Reasonable evidence but interpretation, fuzzy matching, or single-source.
Low	`0.30 – 0.59`	Inferred or weakly evidenced; not suitable for automated blocking decisions.
Very low	`< 0.30`	Internal-only; not exposed via default customer-facing surfaces.

5.1 Confidence decay rules

- Cross-chain propagation reduces confidence by at least 0.2 from the parent label, and may only propagate from parents with confidence `>= 0.8`.
- 2-hop propagation reduces confidence by at least 0.3 from the source.
- Stale labels (`lastVerified` older than 365 days) are subject to verdict downgrade in downstream consumers, regardless of stored confidence (see § 10).

5.2 Mapping to downstream products

Downstream products map the float confidence into product-specific tiers. BA | Oracle, for example, uses four bands (`high`, `medium`, `low`, `none`) and exposes only high-band labels by default; low-band labels require explicit opt-in by enterprise customers (documented in ADR-0001 of the BA repository).

---

6. SANCTIONED — Strict Definition

The category `SANCTIONED` is reserved and used in a strictly defined manner because it carries legal weight for our customers.

6.1 What qualifies

A label may bear the `SANCTIONED` category only when the underlying address is, **on its own merits**, designated on one or more of the following official lists, by direct named or aliased match:

1. OFAC SDN list (United States),
2. EU Consolidated Financial Sanctions List (FSF),
3. UN Security Council Consolidated List,
4. UK HM Treasury OFSI consolidated list, or

5. Swiss SECO SESAM list.

6.2 What does not qualify

- Research-level claims, journalistic attribution, or non-official designations: stored as `category: OTHER` with `tag: unverified-sanctioned`.
- Addresses near (1-hop, 2-hop, multi-hop) a sanctioned address: not exposed as `SANCTIONED`. May appear via inferred-label channels with appropriate tagging and threat-level capping (§ 7).
- Addresses controlled by an entity that is sanctioned but which itself has not been listed: not exposed as `SANCTIONED` unless and until the address itself appears on an official list.
- AI-classifier outputs in the `SANCTIONS_RISK` pattern class: explicitly demoted (§ 8).

6.3 Why this matters

When a customer blocks a transaction on a `SANCTIONED` label, the customer makes a legally consequential decision. Conflating research-level claims with official designations would create both regulatory exposure for the customer (over-blocking sanctioned addresses that aren't actually sanctioned exposes them to civil liability) and reputational exposure for BlockchainAnalysis.io. The strict scope is non-negotiable.

---

7. Inferred Labels Policy

A label is "inferred" when it is derived from analysis rather than observation. Inferred labels are essential for coverage but must be clearly bounded.

7.1 What counts as inferred

A label carries the tag `inferred` and at least one method tag when it is produced by:

- **Clustering** (`clustering` tag) — UTXO co-spend, Solana program account, or other address-grouping heuristics.
- **2-hop propagation** (`2-hop` tag) — risk inherited from a counterparty's counterparty.
- **Cross-chain inference** (`cross-chain` tag) — same-entity attribution across chains via ENS, deposit address, or pattern matching.
- **First-funder analysis** (`first-funder` tag) — inheritance from the address that funded the wallet's first transaction.
- **AI-derived classification** (`ai-derived` tag) — pattern-matching output from an LLM or ML classifier.

7.2 Mandatory caps

Inference type	Threat level cap	Confidence cap	Verdict cap (Oracle)
Clustering	`MEDIUM`	inherits from cluster centroid, ≤ 0.85	`warning`
2-hop propagation	`MEDIUM`	parent − 0.3	`warning`
Cross-chain inference	`MEDIUM`	parent − 0.2; parent must be ≥ 0.8	`warning`
First-funder	`LOW` (informational)	depends on parent attribution	`warning`
AI-derived	`MEDIUM`	classifier confidence, but never used to drive `danger`	`warning`

7.3 Why inferred ≠ direct

Direct attribution means the address itself appears on a list, in a public disclosure, or in observable on-chain evidence. Inference connects an address to a flagged source via a heuristic that, while statistically sound, carries non-trivial false-positive risk. A wallet that received funds from a Tornado Cash output is not, by itself, a money launderer. A wallet clustered with a scam deployer may be a victim. The methodology refuses to elide that distinction.

---

8. AI Classifier Demotion Policy

BA | Intel uses ML and LLM classifiers as one signal among many. Classifier outputs are subject to a strict demotion policy at the normalization layer, even when the upstream classifier reports high confidence.

8.1 The demoted classes

The following classifier output classes are remapped before they reach customer-facing surfaces:

- `MONEY_MULE` → demoted to `OTHER` + `other-flag`, capped at `MEDIUM` threat level.
- `SCAM_OPERATOR` → demoted to `OTHER` + `other-flag`, capped at `MEDIUM`.
- `SANCTIONS_RISK` → demoted to `OTHER` + `other-flag`, **never exposed as `SANCTIONED`**.

8.2 Rationale: pattern ≠ proof

A classifier that recognizes the on-chain pattern typical of a money mule is recognizing a pattern. The pattern is consistent with money muling but is also consistent with the legitimate behavior of certain high-velocity users (gig economy payouts, frequent peer-to-peer trading, custodial sweep operations). Treating the pattern as proof would systematically over-block legitimate users.

For the same reason, AI-derived `SANCTIONS_RISK` output cannot reach the `SANCTIONED` category. The legal threshold for `SANCTIONED` (§ 6) is a name on an official list, not a pattern match.

8.3 Where the AI signal is still useful

Demoted AI signals remain accessible:

- As `OTHER` + `other-flag` warnings, visible to compliance officers reviewing borderline cases.
- As enterprise-tier opt-in signals in BA | Oracle (via `include_low_confidence`), with explicit caveat in the response: "low-confidence signal — for review, not action."

---

9. Mandatory Quality Gates

The following gates are enforced on every label written to the database. Non-compliance is a defect and creates legal risk for customers.

9.1 Required fields

Every label must populate, with non-empty values:

- `source` — identifier of the data origin (e.g., `ofac`, `scamsniffer`, `inferred-clustering`).
- `confidence` — numeric `[0.0, 1.0]`, never null.
- `category` — one of the 17 enum values (§ 3).
- `threatLevel` — one of five levels (§ 4).
- `tags` — array; at minimum the method tag for inferred labels.
- `description` — human-readable rationale; never empty.
- `lastVerified` — timestamp of the most recent verification.

9.2 Insertion semantics

- `ON CONFLICT (address, chain) DO NOTHING` is the only permitted insertion mode for the `entity_labels` table. The first high-quality label wins. Later inserts are discarded.
- Bulk imports affecting more than 1,000 rows require a dry-run that emits the proposed inserts to a review log before any write occurs.

9.3 Category–threat-level consistency

The following combinations are disallowed by validation at write time:

- `SCAM`, `PHISHING`, `EXPLOIT`, `MIXER`, `SANCTIONED` with `threatLevel: SAFE`.
- Any inferred label with `threatLevel > MEDIUM`.
- `SANCTIONED` from a source outside the § 6 whitelist.

9.4 Curated token policy

Token-holder labels are added only for tokens with significant adoption (≥ 10,000 holders) or specific analytical value. The full curated list and inclusion rationale is maintained internally; a public summary is available on request.

---

10. Freshness, Stale-Label Policy & Update Cadence

10.1 Update cadence

Source category	Refresh interval	Mechanism
Sanctions lists (§ 2.1)	4 hours	`sanctions-sync` cron
Enforcement actions (SEC, FCA, etc.)	Daily	Per-source cron job
Industry blacklists (§ 2.2)	Per-source: 1 h to 7 d	Per-source ingestion script
Exchange attribution (§ 2.3)	Weekly review + on-demand	Curated list update
Token-holder extraction (§ 2.4)	Weekly per token	RPC extraction script
Behavioral inference (§ 2.5)	Weekly batch	Heuristic pipeline
Academic datasets (§ 2.6)	On publication	Manual ingestion
Cluster expansion (§ 2.7)	On cluster growth events	Recompute pipeline

The `lastVerified` field on each label records the most recent re-verification or re-ingestion timestamp for that specific label.

10.2 Stale-label policy

Labels whose `lastVerified` exceeds 365 days are considered stale. Stale labels remain in the database (they are still useful as historical record) but are subject to downstream verdict downgrade: in BA | Oracle, a `danger` verdict driven solely by stale labels is downgraded to `warning`, with the response explicitly flagging the staleness so that the customer can request re-verification.

10.3 Sync to customer-facing replica

Labels are written to the hz1 master database and synchronized to the ba-intel read replica every four hours. Customer-facing queries read from the replica. Replication lag is monitored; alerts fire if lag exceeds 30 minutes.

10.4 Retirement

Labels are retired (soft-deleted) when:

- The underlying source officially removes the designation (e.g., an OFAC delisting).

- A re-verification cycle determines the original attribution was incorrect.
- A successful false-positive dispute (§ 11) is upheld.

Retired labels remain available for audit purposes via internal queries; they are not returned to customer-facing queries.

---

11. False Positive Reporting, Disputes & Service Levels

11.1 Reporting channel

False-positive reports may be submitted at any time via the customer dashboard or by emailing `disputes@blockchainanalysis.io`. A report must include:

- The address and chain in question.
- The label as observed (category and source).
- The basis for the dispute (evidence of legitimate identity, evidence of incorrect source, etc.).

11.2 Service level

- Acknowledgement: within 1 business day.
- Initial triage: within 3 business days.
- Resolution: within 5 business days for the typical case; complex cases (multi-source attribution, ongoing investigation) may extend to 10 business days with progress updates.

11.3 Resolution outcomes

- **Upheld** — the label is retired or recategorized; the change propagates to the read replica on the next sync.
- **Modified** — the label remains but its description, confidence, or tags are updated to reflect new information.
- **Rejected** — the label stands; the rationale is recorded and shared with the reporter.

Every dispute outcome is logged in an immutable audit table.

11.4 Customer-facing posture

BA | Intel publicly commits to treat dispute resolution as a first-class operational obligation. The dispute log is auditable by customers under the terms of their DLA.

---

12. Legal Defensibility, Versioning & Change Management

12.1 Per-label traceability

Every label can be traced to its `source`, `lastVerified` timestamp, and `confidence`. There are no labels of unknown origin in the customer-facing dataset. Customers may, under DLA terms, request the full provenance chain for any label.

12.2 Methodology versioning

This methodology document is versioned. The version number applies to the methodology as a whole. Updates are classified as:

- **Patch** (1.0.x) — clarifications, editorial fixes, no semantic change.
- **Minor** (1.x) — additions to source categories, new label categories, new tags, expanded customer workflows — backward compatible.
- **Major** (2.x) — semantic changes that would affect customer integrations: redefining a category, changing the `SANCTIONED` whitelist, modifying the inferred-label cap rules.

12.3 Notice period for breaking changes

Major methodology changes and breaking schema changes affecting customer-facing data are announced with a minimum 90-day notice period to all DLA customers. Notice is delivered via email to the technical and compliance contacts on file, and is also published on the customer dashboard.

12.4 Audit posture

- Annual external audit planned Q4 2026 (SOC 2 Type II engagement).
- Internal quality review quarterly: tier distribution, false-positive rate, source freshness, dispute throughput.
- Change log for the methodology document is maintained in the public repository.

12.5 Suitability for DLA, SOC 2, and GDPR

This document is designed to be attachable to:

- **Data License Agreements** — as the canonical reference for what the licensed data is, how it is produced, and what the customer can expect.
- **SOC 2 Type II evidence packages** — as documentation of CC7 (System Operations) and CC8 (Change Management) controls applied to the dataset.
- **GDPR Article 30 records** — as documentation of the categories of data processed, the source categories, the purposes, and the retention rules. Where customer workflows lead to processing of personal data crosswalked from on-chain identifiers, the customer is the controller and is responsible for the lawful-basis assessment for their own use.

Change log

Version	Date	Notes	
1.0	2026-05-20	Initial public publication. 17 categories (with `STABLECOIN_ISSUER` newly added vs prior internal 16-category taxonomy). Codifies inferred-label caps, AI classifier demotion policy, and stale-label downgrade rules consistent with BA	Oracle ADR-0001.

BlockchainAnalysis.io — operating entity Advisorn GmbH (Switzerland), transitioning to BA UK Ltd / BA DIFC / BA Delaware LLC by jurisdiction. Contact for compliance and DLA inquiries: `contact@blockchainanalysis.io`.