# Millions — Data Provenance & Audit Playbook (v0.1)

Purpose: give you **complete control** over where every field in the Millions dashboards comes from, how it’s computed, when it refreshes, and how to troubleshoot when something looks off. This is a **living doc** we’ll expand as we wire Circa, EPA, injuries, and weather.

## 0) Scope & Goals

* **Traceability**: Every dashboard value maps to an upstream source + script + transform.
* **Auditability**: Weekly checks flag missing/implausible values with specific remediation steps.
* **Repeatability**: One‑click commands to regenerate, validate, and export.

**Key files involved**

* Builder: scripts/millions\_build\_game\_view.py
* Audit: scripts/millions\_audit\_week.py
* Member export: scripts/millions\_export\_member\_html.py
* Inputs: picks/survivor/survivor\_roadmap\_expanded.csv, data/2025\_dvoa\_projections.csv (W1), data/2025\_dvoa\_season.csv (W2+ planned), data/circa\_week\_{W}.csv (planned), data/epa\_team\_s2d.csv (planned)
* Outputs: picks/millions/millions\_weekly\_games.csv, picks/millions/millions\_roadmap\_game.csv, picks/millions/diagnostics/millions\_week\_audit.md

## 1) Column Dictionary (what it means, where it comes from, how it’s made)

Use this as your quick reference when something looks wrong. We’ll expand/lock this as sources go live.

### Keys & Display

| Column | Type | Source | Transform/Logic | Refresh | Owner |
| --- | --- | --- | --- | --- | --- |
| week | int | Survivor expanded CSV | Direct copy | Weekly | builder |
| game\_num | int | Survivor expanded CSV | Direct copy/order | Weekly | builder |
| home\_team, away\_team | str | Survivor expanded CSV | Alias map applied (see registry) | Weekly | builder |
| matchup | str | Derived | "{away} @ {home}" | Weekly | builder |
| kickoff\_local | str | Survivor expanded CSV | Direct copy (Eastern) | Weekly | builder |
| venue | str/NaN | Survivor expanded CSV (if present) | Direct | Weekly | builder |

### Market Lines (Current/Open/Closing/Circa)

| Column(s) | Type | Source | Transform/Logic | Refresh | Owner |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| current\_spread\_home/away | float | Survivor expanded CSV (consensus/current) | Robust fallback + sign‑flip backfill; provenance in current\_spread\_status | Mon–Thu | builder |  |  |  |
| open\_spread\_home/away | float | (Planned) market open feed | Direct; sign relative to home | Weekly | builder |  |  |  |
| closing\_spread\_home/away | float | (Planned) market close feed | Direct; post‑mortem only | Mon (post) | builder |  |  |  |
| circa\_spread\_home/away | float | Circa PDF/CSV (planned) | Parsed Thursday AM; sign normalized to **home** | Thu AM | circa parser |  |  |  |
| circa\_spread\_price\_home/away | float | Circa PDF/CSV (planned) | Price/juice fields if available | Thu AM | circa parser |  |  |  |
| current\_spread\_status | str | Builder | One of `current:primary | secondary | flip | fallback0` | Mon–Thu | builder |

### Line Value (computed only when Circa exists)

| Column(s) | Type | Source | Transform/Logic | Refresh | Owner |
| --- | --- | --- | --- | --- | --- |
| line\_value\_home/away | float | Derived | current - circa (aligned to same side); hidden on Member pre‑Thu | Thu+ | builder |
| line\_value\_current\_home/away | float | Derived | Same as above (explicit current vs Circa) | Thu+ | builder |
| line\_value\_closing\_home/away | float | Derived | closing - circa (post‑mortem) | Mon (post) | builder |

### Team Quality & Context

| Column(s) | Type | Source | Transform/Logic | Refresh | Owner |
| --- | --- | --- | --- | --- | --- |
| total\_dvoa\_\* / off\_dvoa\_\* / def\_dvoa\_\* / st\_dvoa\_\* | float/str | W1: data/2025\_dvoa\_projections.csv; W2+: data/2025\_dvoa\_season.csv | Join by team with alias map; display strings allowed | W1 (static) / Weekly | builder |
| off\_epa\_per\_play\_\*, def\_epa\_per\_play\_\* | float | data/epa\_team\_s2d.csv (planned) | Join by team; sanity range ≈ [−0.3, +0.3] | Wed+ | builder |
| epa\_diff\_home/away | float | Derived (planned) | Offense of side − opponent defense | Wed+ | builder |
| rest\_days\_home/away | int | Survivor expanded CSV | Direct copy | Weekly | builder |
| rest\_days\_diff | int | Derived | rest\_home − rest\_away | Weekly | builder |
| injuries\_key\_home/away | str | Snapshot feed (planned) | Short roll‑up notes: QB (Q), LT (D) | Tue/Thu | injuries ingest |
| weather\_notes | str | Snapshot feed (planned) | e.g., Rain 60%, 12 mph or INDOOR | Thu/Sat | weather ingest |

## 2) Provenance & Status Flags

* ``: marks how we sourced/filled current lines.
  + current:primary (ideal), current:secondary, current:flip (sign corrected), fallback0 (last‑resort).
* **(Planned)** `:parsed:pdf,parsed:csv,manual`.
* **(Planned)** \*\*/\*\***/**`:fresh:YYYY‑MM‑DD hh:mmorstale:N days`.

## 3) Weekly Audit Checklist (automated + manual)

**Automated via** `` (to extend):

* Missingness thresholds:
  + current\_spread\_\* ≤ 5% missing pre‑Thu; **PASS/ATTN** label.
  + DVOA 0% missing; alias misses listed.
  + EPA/injuries/weather: explicitly labeled **Expected pre‑Thu** until wired.
* Range checks:
  + DVOA within approx. ±40%.
  + EPA within approx. −0.3..+0.3.
  + Spreads within −25..+25.
* Consistency checks:
  + matchup ↔ home\_team/away\_team join sanity.
  + rest\_days\_diff = rest\_home − rest\_away.

**Manual spot‑check (3–5 games):**

* Cross‑verify current line vs public board.
* Confirm DVOA magnitudes & signs.
* If Circa is present (Thu+), inspect line\_value\_\* math on 1–2 games.

## 4) Troubleshooting Playbook (by symptom)

* **Current spread missing or** ``**:**
  1. Check survivor CSV columns (current\_spread, consensus\_spread names).
  2. Confirm alias map hit for both teams.
  3. Re‑run builder; inspect current\_spread\_status.
* **DVOA blank for a team:**
  1. Open DVOA CSV; ensure alias exists (e.g., WAS → WSH, NWE → NE).
  2. Update alias registry; re‑run builder.
* **EPA wildly out of range:**
  1. Confirm unit (per‑play) and season filter.
  2. Check join keys + alias; re‑export.
* ``\*\* looks inverted:\*\*
  1. Ensure both Circa and Current use **home‑relative** signs.
  2. Verify sign‑flip logic for away rows.
* **Injuries/weather empty after snapshot:**
  1. Confirm feed files exist for the week.
  2. Verify ingest script wrote to columns expected by builder.

## 5) Command Cadence (single source of truth)

# From project root  
.\venv\Scripts\Activate.ps1  
  
# 1) Build  
python -m scripts.millions\_build\_game\_view --week 1  
  
# 2) Audit (adds expected-pre‑Thu labels)  
python -m scripts.millions\_audit\_week --week 1 --write-report  
  
# 3) Export Member (unbiased)  
python -m scripts.millions\_export\_member\_html --week 1 --out picks/millions/member\_week\_1.html

## 6) Roadmap to Full Provenance

* Add `` to builder once the parser lands.
* Add ``\*\* timestamps\*\* for EPA/injuries/weather snapshots.
* Add **schema validator** (utils\_schema.py) for required columns + types; fail fast before export.
* Emit a **Provenance Appendix** section at bottom of millions\_week\_audit.md (per column: source → script → transform → last updated).

## 7) Alias Registry (single source of truth)

Keep in a tiny util (importable in builder + joins). Extend on first miss.

{  
 'WAS':'WSH','JAC':'JAX','NOR':'NO','TAM':'TB','SFO':'SF','ARZ':'ARI',  
 'GNB':'GB','KAN':'KC','NWE':'NE','SDG':'LAC','STL':'LAR','OAK':'LV',  
 'Washington':'WSH','Jacksonville':'JAX','New Orleans':'NO','Tampa Bay':'TB',  
 'San Francisco':'SF','Arizona':'ARI','Green Bay':'GB','Kansas City':'KC',  
 'New England':'NE','San Diego':'LAC','St. Louis':'LAR','Oakland':'LV'  
}

## 8) Add‑On: Column Spec Template (for any new field)

Copy this block into the table above when you add a field:

Name: <column\_name>  
Type: <int/float/str/datetime>  
Source: <file or feed>  
Transform: <join keys, math, unit assumptions>  
Refresh cadence: <Mon/Tue/Thu…>  
Owner script: <which script writes it>  
Audit checks: <null threshold, range guards>  
Troubleshooting: <likely failure points + fixes>

### Summary

This playbook keeps our **Member** view clean and unbiased while giving your **Master** dashboard iron‑clad lineage. As we plug in Circa/EPA/Injuries/Weather, we’ll append concrete sources and lock the audit thresholds so you immediately see where anything drifted.