Data Sourcing & Extraction Methodology

Alternative Investment Commitment Data — U.S. Public Pension Funds

Overview

This dataset captures private equity and alternative investment commitments disclosed by five major U.S. public pension systems, representing over $1.5 trillion in combined AUM. The pipeline extracts structured data directly from official government disclosures — no web scraping of third-party aggregators, no LLM-generated values, no manual data entry.

2,125 commitment records across **1,610 unique funds**, with **345 funds independently verified across 2+ pension systems.**

Data Sources

CalPERS — California Public Employees' Retirement System

Source	calpers.ca.gov — PEP Fund Performance (printer-friendly page)
Format	HTML table, parsed deterministically with BeautifulSoup
As-of date	March 31, 2025
Records	429 commitments
Fields	Fund name, vintage year, commitment, capital called, distributed, remaining value, net IRR, net multiple
Confidence	1.0 — structured HTML, zero parsing ambiguity

CalSTRS — California State Teachers' Retirement System

Source	calstrs.com — Private Equity Portfolio Performance Table
Format	PDF, parsed with pdfplumber word-level extraction
As-of date	June 30, 2025
Records	473 commitments
Fields	Fund name, vintage year, commitment, contributed, distributed, market value, net IRR, net multiple
Confidence	0.95 — structured PDF table with consistent column layout

WSIB — Washington State Investment Board

Source	sib.wa.gov — Quarterly Private Equity IRR Report (Q2 2025)
Format	PDF, parsed with pdfplumber word-level extraction
As-of date	June 30, 2025
Records	462 commitments
Fields	Fund name, commitment date, commitment, paid-in, unfunded, distributions, market value, total value, net multiple, net IRR
Confidence	0.90 — structured PDF; minor header artifacts, data rows parse cleanly

Oregon PERS — Oregon Public Employees Retirement System

Source	oregon.gov/treasury — OPERF Private Equity Portfolio (Q3 2025)
Format	PDF, parsed with pdfplumber word-level extraction
As-of date	September 30, 2025
Records	402 commitments
Fields	Fund name, vintage year, commitment, contributed, distributed, fair market value, total value multiple, net IRR
Confidence	0.95 — structured PDF table with dollar-denominated columns

NY Common — New York State Common Retirement Fund

Source	osc.ny.gov — Annual Asset Listing 2024 (pages 170–175)
Format	PDF, parsed with pdfplumber word-level extraction
As-of date	March 31, 2024
Records	359 commitments
Fields	Fund name, date committed, commitment, contributed, distributions, fair value, total value, net multiple
Confidence	0.95 — structured PDF table; IRR not disclosed by this source

Extraction Approach

Deterministic Parsing — No LLM Dependency

Every record in this dataset was extracted using deterministic, rule-based parsing. No large language model was used to read, interpret, or generate any data values.

HTML sources (CalPERS): Parsed with BeautifulSoup. Each table row maps directly to a commitment record. Field mapping is explicit and verifiable.

PDF sources (CalSTRS, WSIB, Oregon, NY Common): Parsed with pdfplumber's word-level extraction. Words are grouped into rows by y-coordinate proximity, then assigned to columns by x-position against source-specific column boundaries calibrated to each document's layout.

Every record carries its source URL, document name, extraction method, and a confidence score. The pipeline is fully reproducible — running it twice on the same source documents yields identical results.

Field Completeness

Field	Coverage
Commitment amount	99.9%
Vintage year	99.7%
Capital called	98.4%
Net multiple	98.0%
Capital distributed	94.8%
Net IRR	52.6% (4 of 5 sources disclose; NY Common does not)
Remaining value	84.8%

Entity Resolution

When the same private equity fund appears in multiple pension portfolios — often under slightly different names — the pipeline links those records to a single canonical fund entity.

Method: Multi-signal fuzzy matching using the rapidfuzz library. A match requires **at least two** of the following signals to agree:

1. **Name similarity** — Token-sort ratio above 85% between normalized fund names
2. **General partner match** — GP name similarity above 85%
3. **Vintage year match** — Exact year agreement

Additional safeguards prevent false positives: fund sequence numbers (e.g., “Fund III” vs. “Fund IV”) must match exactly, strategy keywords (credit, Asia, Europe, infrastructure) must agree, and a minimum distinctiveness threshold filters out matches on generic terms.

Result: 345 funds are independently cross-linked across two or more pension systems. All matching decisions are logged with the raw input name, resolved fund ID, match type, and similarity score for full auditability.

Cross-System Validation

Where the same fund appears across multiple pension portfolios, reported performance metrics serve as an independent consistency check.

Net IRR agreement: For funds reported by two or more systems with overlapping reporting periods, 95% of paired IRR values agree within 2 percentage points — consistent with expected differences in reporting date and fee structures.
Vintage year agreement: 100% exact match across all cross-linked funds.
Commitment reasonableness: All values validated against expected ranges ($1M–$5B per commitment, multiples 0.5x–4.0x, IRRs -20% to +50%). Outliers flagged for manual review.

Provenance & Reproducibility

Every commitment record in the database carries:

Source URL pointing to the original government disclosure
Document name and page number (for PDFs)
Extraction method (deterministic_html or deterministic_pdf)
Extraction confidence score (0.90–1.0)
Extraction timestamp and as-of reporting date

The pipeline is idempotent — re-running against the same source documents produces identical output with no duplicate records. Each run is logged with records extracted, updated, and flagged.

Pipeline built in Python. Data stored in SQLite with full relational schema. Exports available as CSV or direct database query.