Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

Private Equity Data Provider Lineage: A Field-Level Test

> Build a more reviewable acquisition pipeline with Docket. Our deal-origination software and managed research help private equity teams screen companies against a mandate and understand the evidence behind each finding.

Published by Docket · 2026-11-17 · private equity data provider lineage · data lineage · source independence · private company data providers · provenance · corroboration · pe research stack · deal sourcing data · origin graph · due diligence

Original article: https://docket.capital/articles/private-equity-data-provider-lineage-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes in How Agreement Should Be Treated
4. What Public Documentation Reveals, and What It Cannot
5. The Field-Level Lineage Test
6. Worked Example with Fictional Providers (Hypothetical Example)
7. Marginal Value Beyond Corroboration
8. Data Analysis and Evidence
9. Case Studies and Real-World Examples
10. Implementation Considerations and Process Changes
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

Two providers returning the same ownership value is a fact about agreement, not about evidence. For a research or audit team, the question is whether the second subscription adds a **new evidence origin** for a specific field, adds only **distribution or access** to an origin already held, or leaves independence **unknown**. This report proposes a field-level source-lineage test and a marginal-corroboration worksheet to answer that question, and it works the method on wholly fictional records. No vendor was measured, and no vendor's dependence on another is asserted.

The test rests on three separations. First, provider brand, returned URL, publisher, underlying document and original observation are different layers. Provenance standards already model this: the W3C PROV Primer treats a quotation as one entity quoted from another (www.w3.org), and each revision as a new entity (www.w3.org). Second, dependence is claim-specific and time-specific; the 2009 PVLDB copy-detection paper notes that a source can stop copying and become independent (www.vldb.org). Third, a collection timestamp is not a fact date: OpenCorporates documents its provenance date as when it retrieved the data or made the match (api.opencorporates.com).

Public documentation supports the need for the test but cannot replace it. PrivCo states that multiple independent sources let every material data point be corroborated (www.privco.com), while Altss says source attribution is visible for most fields, not all (altss.com). Those are vendor statements. Several vendors also describe shared upstream families, such as regulatory filings and news (preqin.com), so a shared origin is plausible even when no copying occurs.

Agreement can be misleading in measurable ways. A 2009 experiment found naive voting concluded that only 1192 of 5269 restaurants had ever existed (www.vldb.org), and in a later stock-data study, 11 of 55 sources (20%, computed here) formed one group derived from a single market-data service (arxiv.org). Neither domain

is company data. Authoritative single records can also be sufficient: the SEC describes Form D as a notice of an exempt offering of securities that is required to be filed with the SEC under Regulation D (www.sec.gov), which fits legal-status facts but not necessarily a current ownership claim.

The recommended acceptance rule is simple. Count **known origin groups** per claim as a lower bound, never as the total number of independent sources. Do not let a majority vote overwrite a material unknown. Price access, freshness and entity-matching benefits separately from corroboration. A shared origin on one copied field is not a reason to cancel a subscription.

Introduction and Background

Private equity (PE) origination and audit teams commonly license more than one company-data platform. The usual reason is coverage: each platform surfaces different targets. A second reason emerges during diligence: when two platforms agree that a target is founder-owned, or that it fits a mandate, teams treat the agreement as corroboration. That inference is the subject of this report.

The inference fails when the platforms share an upstream origin. Provider markets overlap heavily: Opimas, a research firm, identified about 20 leading vendors of private equity data (www.opimas.com) and estimated the market at more than US$1.2 billion in 2022 (www.opimas.com). Vendors describe overlapping source families, including regulatory filings, press releases and news (preqin.com), public and self-reported information (pitchbook.com), and contributions from investors (support.crunchbase.com). Overlapping families do not prove overlapping observations, but they make independence something to test rather than assume.

This report distinguishes four ideas that procurement conversations often merge:

- **Agreement:** two outputs carry the same normalized value.
- **Truth:** the value matches the real-world state at the relevant date.
- **Authority:** the origin is the body best placed to state the fact, for example a registry for legal status.
- **Independence:** the observations were generated by separate processes, so errors in one do not propagate to the other.

Agreement does not imply truth, and independence does not imply authority. The report concerns the fourth idea, applied to the decision of whether an additional subscription supplies a new evidence origin for a given field.

The scope is deliberately narrow. It is not a vendor ranking, a ledger for resolving conflicting values, or an AI benchmark. It addresses agreeing values. The publisher, Docket, offers sourced target research through Triage, Scout and Audit agents and is therefore a direct provider in the broader category; for that reason it is not tabulated below and no performance claim is made about it; buyers should apply the same ledger to its outputs.

Evidence limits are stated up front. No controlled multi-provider trial was run for this report. No vendor export was purchased or inspected. Every vendor description below is attributed to the vendor as a vendor statement, and the vendor pages were checked in November 2026. The worked example uses synthetic providers and values, labeled as such, and the numbers in it are not vendor benchmarks.

Key Changes in How Agreement Should Be Treated

The following five changes move a research stack from counting providers to counting evidence origins.

Change 1: Separate agreement, truth, authority and independence

A single authoritative record can be the right evidence for some claims. The SEC states that a company conducting a Regulation D offering must have an EDGAR account (www.sec.gov), and that the filed Form D becomes public on EDGAR (www.sec.gov). For the claim "this issuer filed a Form D on this date," that filing is sufficient evidence, and demanding a second source adds cost without adding support.

The same filing is weaker for other claims. The SEC's Form D data-set page states that the data are derived from information provided by individual registrants and that the agency cannot guarantee their accuracy (www.sec.gov). A self-reported filing is an authoritative origin for what the issuer reported on the filing date, not a verified account of present control. The acceptance rule therefore depends on the claim, not on the provider count.

Change 2: The unit of analysis is the field, not the provider

Dependence is not a property of two vendors in the abstract. The foundational copy-detection work states that sharing common data does not in itself imply copying (lunadong.com). Two providers can share a registry origin for legal status, differ in origin for headcount, and be unknowable for revenue. Each provider and field pair needs its own disposition.

This also means a test result should not be generalized to a vendor name. A finding on "registered legal name" says nothing about "current majority owner." The ledger described later records field and period explicitly for this reason.

Change 3: Separate five layers between a screen value and the world

A value displayed on a platform passes through up to five layers, and independence can be lost or preserved at any of them:

- **Provider brand:** the platform or profile ID the analyst sees.
- **Returned URL:** the link the platform exposes as its source.
- **Publisher identity:** the organization operating that URL.
- **Underlying document:** the specific filing, release or page containing the claim.
- **Original observation:** the act of generating the fact, such as a registrant reporting an owner.

Provider metadata may stop at different layers. OpenCorporates documents a source URL as the URL from which the data was obtained (api.opencorporates.com) and a separate publisher field as the original publisher of the information (api.opencorporates.com). The documentation also describes a provenance type for data reconciled to a company by OpenCorporates itself (api.opencorporates.com), which is a derived layer, not an observation. The report examined documentation only, not exports, so no claim is made about how any field is populated in practice.

Change 4: Keep three dates apart

Three different dates are routinely merged. The **fact date** is when the claimed state was true. The **publication date** is when a document or registry entry appeared. The **retrieval date** is when a provider or analyst collected it. The Open Ownership Beneficial Ownership Data Standard separates these: its statement date is when the information is claimed to be true (standard.openownership.org), its publication date is when the statement was published (standard.openownership.org), and its retrieval field applies only where data is republished (standard.openownership.org).

Registries show why this matters. UK guidance asks companies to report changes in people with significant control within 14 days of confirming the change (www.gov.uk), so the register date reflects the company's confirmation, not necessarily the date control changed. A provider that stamps its own retrieval date on the same entry has added a third date, not a new observation.

Change 5: A majority vote is not corroboration

The Dong, Berti-Equille and Srivastava paper illustrates the failure mode with a five-source example in which copies form the majority, so that naive voting would make wrong decisions for three researchers (lunadong.com). A survey of truth discovery notes the structural reason: majority voting has no chance to select the correct value when most sources claim the same wrong one (arxiv.org).

The same survey summarizes the usual detection heuristic, that sources making many common mistakes are not likely to be independent (arxiv.org), and states its limit: the principle becomes ineffective when a source copies a good source (arxiv.org). That limit matters for procurement. If every agreeing value is correct, there are no shared errors to inspect, so agreement on clean data cannot distinguish independent origins from copies. Only recorded lineage can.

What Public Documentation Reveals, and What It Cannot

Vendor sourcing pages are the first lineage input available to buyers, and they vary in specificity. *Table 1* below summarizes what selected providers state about their own sourcing. All entries are vendor statements retrieved in November 2026; none was verified against exports, and none shows that one provider depends on another.

Provider	Stated source families (vendor claim)	Per-field attribution stated?	What the statement cannot establish
PrivCo	Aggregates signals from dozens of third-party providers (www.privco.com); licenses datasets for firmographics (www.privco.com); labels estimates as estimates	Not stated on the page reviewed	Which licensed providers, and whether any field's "independent sources" share an upstream
Altss	Labels the source type behind each value (altss.com)	Stated for most fields; deeper audits offered to enterprise clients (altss.com)	Which fields lack attribution; a missing verification date is treated as unknown (altss.com)

Provider	Stated source families (vendor claim)	Per-field attribution stated?	What the statement cannot establish
PitchBook	Public and self-reported data (pitchbook.com); web crawlers (pitchbook.com); direct calls and emails (pitchbook.com)	Not stated on the pages reviewed	Which origin produced a given field value
Crunchbase	Venture partners and contributors rather than web scraping (support.crunchbase.com); algorithms and government filings (support.crunchbase.com)	Not stated on the pages reviewed	Whether contributor-supplied fields repeat what contributors supply elsewhere
Preqin	Contributions from investors and service providers (preqin.com); filings, releases, news and websites (preqin.com)	Not stated on the pages reviewed	Field-level split between contributed and public sources
Sourcescrub	Categorizes each company and the source where information is found (www.sourcescrub.com)	Source-to-company mapping described; per-field attribution not stated	Whether mapped sources are distinct publishers or copies
Grata	Combines AI, human research and non-public insights (grata.com)	Not stated on the page reviewed	Which observations come from which component

The table shows three patterns. Field-level attribution is explicitly claimed by one of the seven third-party rows, and only for most fields. Several providers name upstream families that overlap by type (filings, news, contributed data), though type overlap is not instance overlap. And every row has a gap between a collection statement and a per-field origin. Absence of a statement on a page does not show the capability is missing; the buyer must request the metadata in a trial.

Two further points from the documents. PitchBook says it does not engage in data-for-data exchanges (pitchbook.com), a relevant disclosure because exchange arrangements are one way a value could travel between providers; the statement is a vendor claim. Crunchbase says third-party datasets are available on its platform (support.crunchbase.com), which means a single screen can mix first-party and licensed fields, a reason to record the origin per field rather than per platform.

Academic work on commercial databases supports the same caution. Kaplan and Lerner write that existing databases differ in methodologies and that analyses frequently produce discrepancies (www.nber.org). Harris, Jenkinson and Kaplan state that each of four fund-performance datasets gathers information differently (www.nber.org), and then reason that agreement despite very different data collection and reporting methods is informative (www.nber.org). That last argument is the independence logic this report formalizes: agreement counts for more when collection methods are documented to differ.

The Field-Level Lineage Test

The test has five steps. It is designed to be run on a permitted trial sample, using only data a buyer is licensed to retain, and it stores origin metadata rather than private source contents. Rights to retain and reuse metadata and excerpts should be confirmed against each contract before a trial; that licensing analysis is outside this report.

Step 1: Build a field-origin ledger

For each inspected observation, record one ledger row. The ledger is a worksheet, not a database, and a spreadsheet suffices for a trial of a few dozen fields. Suggested column groups:

- **Identity:** provider and profile ID; canonical entity; field and period.
- **Value:** raw and normalized value; estimate status.
- **Origin:** returned URL; original publisher or document ID; explicit attribution.
- **Dates:** source date; retrieval date.
- **Derivation:** transformation applied by the provider.
- **Judgment:** origin-group ID; independence disposition; supporting excerpt; reviewer.

Estimate status needs its own column because estimates are not observations. Altss defines an estimated label for values inferred where no source states the value (altss.com), and PrivCo says it marks estimates as estimates (www.privco.com). Two providers returning similar estimated headcounts are two model outputs, and their agreement says nothing about an observed fact unless the model inputs are known to differ.

Step 2: Draw a claim-level origin graph

For each claim, draw a directed graph: returned profile, then document or publisher, then original observation. Label each edge as direct, quoted, derived, inferred or unknown. The W3C PROV vocabulary supplies the labels. PROV defines derivation as one entity's existence and characteristics being at least partly due to another (www.w3.org), and the PROV data model defines a revision as a derivation yielding a revised version of an original (www.w3.org).

A graph records claims about dependence, not proof of truth. An edge marked unknown stays unknown; it is never silently promoted to independent. For each edge, store the evidence that supports the label, for example an explicit attribution link, a matching filing number, or a documented syndication relationship.

Step 3: Recognize known dependencies and refuse weak inferences

Known dependencies include explicit attribution, matching source or filing identifiers, and acknowledged syndication. Syndication is a documented mechanism, not an exotic one. PR Newswire describes displaying releases on thousands of syndicated websites (prnewswire.com), and Google's guidance on syndicated content notes that the most effective way to avoid duplication is for partners to block indexing of the content (developers.google.com). Three web pages carrying the same release are one release.

Weak signals do not establish dependence or independence. Identical values, similar wording, shared hosting and corporate affiliation can each occur between independent observers or between copiers. The copy-detection literature makes the asymmetric point that sharing the same false value is typically a rare event (lunadong.com), which is why shared errors are informative while shared correct values are not.

Step 4: Classify each provider and field pair

Table 2 below defines four dispositions. The labels deliberately avoid the word "independent" for anything that has not been documented.

Disposition	Evidence required	What the second provider adds	Decision implication
Same documented origin	Matching document ID, filing number or explicit attribution to the same publisher	Access, normalization or freshness only	Count one origin group; price convenience separately
Independently documented origins	Each value traces to a distinct document and observer, with edges recorded	A new evidence origin for that field	Count two origin groups; document the evidence for each edge
Partially shared derivation	One value is derived or quoted from, or revises, the other's origin, with other inputs distinct	Possibly an additional input, not a full second origin	Count shared portion once; list distinct inputs
Opaque or unknown	No metadata, an estimate without inputs, or an undisclosed licence	Nothing that can be credited	Record as unresolved; do not credit as corroboration

The table implies an asymmetry in burden of proof. Independence needs positive documentation; dependence needs only one recorded edge; unknown needs nothing and is the default. Because most vendor pages do not state per-field origin, unknown is likely to be the most common disposition in a first trial, and that is a finding, not a failure of the method.

The date columns need the same discipline. A **time-lag** pattern, where one provider shows an older version of a filing than the other, is a revision of one origin, not two. It belongs under the same origin group with a revision edge. Disagreement of this kind is outside this report's scope, which is agreeing values, but it shows why dates are recorded: otherwise a lagging copy looks like a conflicting source.

Step 5: Compute four rates with denominators shown

Four rates summarize a trial. Each is reported with its numerator and denominator and none is turned into a statistical confidence.

- **Agreement rate:** matched comparable fields where values agree, divided by matched comparable fields.
- **Documented-origin coverage:** inspected observations with a recorded origin edge, divided by inspected observations.
- **Known origin groups per claim:** the count of distinct groups documented for that claim, reported as a lower bound.
- **Unknown-lineage share:** inspected observations in the opaque disposition, divided by inspected observations.

A fifth quantity, the **marginal corroboration rate**, is fields where the additional provider contributes documented independent support, divided by inspected comparable fields. It should state whether the supporting origin was sufficient for the claim type. No correlation coefficient or confidence boost should be computed from source counts without a suitable dataset; the 2009 and later truth-discovery studies model dependence with data on known truths, which a procurement trial lacks.

Worked Example with Fictional Providers (Hypothetical Example)

This section is a **Hypothetical Example**. Every provider, entity and value is synthetic. The numbers illustrate arithmetic and are not vendor benchmarks, and nothing here describes any real company or product.

Suppose a team trials **Provider Alpha** (its incumbent) and **Provider Beta** (a candidate) on 20 fictional targets for one field, "current majority owner." *Table 3* below walks through five illustrative scenarios involving a fictional company, **Northgate Components (synthetic)**, and a third fictional source, **Provider Gamma**.

Scenario (all synthetic)	What each provider returns	Origin groups documented	Disposition	Credit as new evidence origin?
A. Three copies of one filing	Alpha, Beta and Gamma each show "Fund X holds 62%" and each returned URL resolves to the same fictional registry filing number	1	Same documented origin	No: access only
B. Press-release syndication family	Beta and Gamma show "Fund X acquired a majority stake" with URLs on three different news sites; all three reproduce one fictional release	1 (the release)	Same documented origin	No: one assertion, three copies
C. Two independent observations	Alpha cites a registry filing by the issuer; Beta cites a separate fictional regulator notice filed by the investor	2	Independently documented origins	Yes, for this claim and field
D. Opaque estimate	Alpha shows "Fund X, est. 55 to 65%" with no inputs; Beta shows "Fund X, est. 60%" with no inputs	Unknown (0 documented)	Opaque or unknown	No: unresolved
E. Time-lag update	Alpha shows the 2026 filing; Beta shows the 2025 filing for the same entity	1 (two revisions)	Same origin, revised version	No: Beta is behind, not a second origin

Scenarios A and B show the failure of counting providers: three agreeing providers amount to one origin group, and their agreement is guaranteed by construction. Scenario C is the only case where the second provider contributes a new origin, and even there independence is documented, not inferred from matching values. Scenario D remains unresolved until inputs are known. Scenario E is a revision within one group, priced as a freshness difference.

Now compute the four rates on the full synthetic sample of 20 matched fields. Suppose the dispositions are: 8 same documented origin, 3 independently documented, 2 partially shared derivation, and 7 opaque. Suppose the two providers' values agree on 15 of 20 fields: all 8 same-origin fields, all 3 independent fields, 1 of the 2 partially shared fields, and 3 of the 7 opaque fields.

- **Agreement rate:** 15 of 20 = **75%**.
- **Documented-origin coverage** for Beta: 8 + 3 + 2 = 13 of 20 = **65%**.
- **Unknown-lineage share:** 7 of 20 = **35%**.
- **Marginal corroboration rate:** 3 of 20 = **15%**, reported as a lower bound because some opaque fields may in fact be independent.

A headline of "75% agreement" would suggest strong corroboration. The lineage view says that, in this synthetic sample, 8 of the 15 agreements are guaranteed by a shared origin, 3 are documented independent support, 1 reflects partial overlap and 3 are agreements between opaque estimates that cannot be credited. The correct statement is: Beta adds documented independent support on at least 15% of inspected comparable fields, and leaves independence unknown on 35%.

Marginal Value Beyond Corroboration

A provider that adds no independent origin for a field can still be worth its subscription. The worksheet below separates the benefits so that they are priced and judged separately. It contains no prices, and the article does not recommend canceling a product based on a copied field.

- **Independent corroboration:** count fields with documented independent support, by claim type, using the marginal corroboration rate above.
- **Access:** note fields where Beta exposes a registry record that Alpha only summarizes; access to the same origin has value for retrieval speed and for analyst checking.
- **Normalization:** note whether Beta resolves entity names, currencies or share classes more consistently. GLEIF's Level 1 reference data answers "who is who" (www.gleif.org), which is the kind of identifier layer that makes cross-provider matching possible.
- **Freshness:** record publication and retrieval dates and measure the lag against the origin. PitchBook says it publishes in six cycles per day (pitchbook.com), a claim a buyer can test by comparing retrieval dates.
- **Entity matching:** record how many entities each provider could map to the canonical entity. Phalippou's review of fund datasets reports that only about half of funds could be matched by name across two providers (ludovicphalippou.substack.com), which shows why matching effort can dominate comparison work.
- **Unknown lineage:** record the share of fields with no metadata and whether the provider will supply origin data on request, which some providers say they offer for sensitive fields.

The output is a column per benefit, not a single score. A provider that scores zero on corroboration and high on normalization is a distribution purchase and should be recorded as one.

Data Analysis and Evidence

The evidence base is thin on direct measurement of company-data providers and richer on adjacent domains. *This section separates what each study measured from what it did not.* Computed ratios are marked as computed.

Copying and voting (non-company domains). In the 2009 PVLDB study, the authors collected weekly data on Manhattan restaurants from 12 web sources (www.vldb.org), detected copying between 12 of 66 source pairs (www.vldb.org) (about 18%, computed), and found naive voting concluded that only 1192 of 5269 restaurants had ever existed (www.vldb.org). The authors state that the true copying relationships were not known, so the detection is a model output. In the deep-web study of stock and flight sources, one stock group had 11 sources with the same layout, schema and highly similar data (arxiv.org); the 55-source stock sample

means that single group held 20% of sources (computed). Removing copiers in the flight data raised the precision of dominant values from .864 to .927 (arxiv.org). None of these figures transfers to company providers.

Database coverage and construction (private capital). Retterath and Braun's benchmark compared eight databases against verified data on 339 financing rounds from 396 venture partnerships (medium.com). The author's summary concludes that a best-coverage composite would draw on different databases for different fields (medium.com), which supports field-level analysis. The study compared databases to ground truth and did not measure database-to-database copying. The OECD's study of Crunchbase describes two channels, an investor network and community contributors (www.oecd.org), and notes that the benchmark used for comparison is typically compiled by national associations, often with the support of commercial data providers (www.oecd.org). A yardstick built partly from commercial data is not a fully independent yardstick.

Imputation (company data). A Federal Reserve paper on the National Establishment Time Series documents that employment is imputed for more than two thirds of establishments with fewer than five employees (www.federalreserve.gov), about 90 percent of firms aged zero or one have imputed employment (www.federalreserve.gov), and sales imputation is around 80 percent among small firms (www.federalreserve.gov). These are documented features of one derived dataset in 2014, and they show why an estimate-status column matters: an imputed value is a model output, not an observation.

Filing mechanics (authoritative records). Form D is due within 15 days after the first sale of securities (www.sec.gov), and the date of first sale is when the first investor is irrevocably contractually committed (www.sec.gov). Companies House states that it lacks the statutory power or capability to verify the accuracy of information companies send it (resources.companieshouse.gov.uk). GLEIF's Level 2 data record the direct accounting consolidating parent reported by entities (www.gleif.org), and a non-consolidating owner is handled as an exception (www.gleif.org). A registry entry is therefore an origin of what was filed or reported, on the filing's own date and definition.

Retrieval and syndication (an analogy from news). A 2025 Tow Center study tested eight generative search tools (www.cjr.org) with 1,600 queries (www.cjr.org) and reported that tools sometimes pointed to syndicated versions of articles (www.cjr.org). It concerns news retrieval, not ownership research, and measures neither current model versions nor company-data error rates.

Case Studies and Real-World Examples

A vendor anecdote that travels (DealRoom and Grata). A DealRoom article on AI in mergers and acquisitions, showing a last-updated date of October 29, 2026 at retrieval, attributes a deal-sourcing anecdote to Grata (dealroom.net). Grata's own article, the apparent origin, carries the same anecdote (grata.com). In origin-graph terms this is a quotation edge from a vendor article to a secondary article. Two pages describe one reported experience; they are not two audited studies. This report omits the individuals and outcomes the anecdote names.

One study, two reports (an analogy: Tow Center and Nieman Lab). Nieman Lab's coverage of the Tow Center work directs readers to the full study at the Columbia Journalism Review (www.niemanlab.org). Counting both articles as two data points would double-count one experiment, which is the same error as

counting syndicated copies.

Common upstream feeds in a derived dataset (NETS). The Federal Reserve paper describes a dataset derived from Dun and Bradstreet establishment records, in which a research firm receives an annual snapshot and links it over time (www.federalreserve.gov). A buyer comparing a derived dataset with its upstream would find high agreement by construction, which is the shared-origin pattern in Scenario A.

Implementation Considerations and Process Changes

Run a bounded trial

Select a small, permitted sample, for example 20 to 40 entities and one to three high-consequence fields. Request the following in writing for each value, and record what the provider declines to supply:

- **Returned URL and publisher:** the link and the organization behind it.
- **Document ID:** the filing, release or page identifier.
- **Source date:** the fact or publication date, kept apart from retrieval date.
- **Estimate flag and transformation note:** whether the value is observed, derived or modeled. Where a provider states that most fields carry visible attribution, a trial tests that claim on your sample and for your fields.

Assign claim-specific sufficiency rules

Not every claim needs two origins. Use the claim type to set a rule: legal status may be sufficient from one authoritative filing; a current ownership percentage or a mandate-fit disqualifier may need support from a source with a documented observer other than the entity itself. FATF's guidance on beneficial ownership describes a multi-pronged approach combining information from several places, including companies themselves (www.fatf-gafi.org), and its page states that countries using it were more effective than those using a single approach (www.fatf-gafi.org). That finding concerns national transparency systems, not vendors, but it supports a rule that high-consequence ownership claims should not rest on a single self-reported source.

Apply an acceptance gate to consequential claims

For ownership or disqualifier claims, a second reviewer should assign one of three outcomes:

- **Sufficient authoritative support:** a documented origin is appropriate for the claim type and its date is current enough for the decision.
- **Separately corroborated:** at least two origin groups are documented with recorded edges, and the observers are distinct.
- **Unresolved:** lineage is unknown, estimates lack inputs, or dates do not align. The claim is flagged and not used to disqualify or confirm a target.

A majority vote never converts an unresolved claim into a supported one. The 2009 paper's statement that straightforward ways to resolve conflicts, such as voting, fall short in the presence of copying (www.vldb.org) is the reason.

Watch identity and date semantics in registries

Registry content has its own limits. Companies House states that basic checks cover completion and signature (resources.companieshouse.gov.uk), and UK identity verification for directors and persons with significant control became a legal requirement on 18 November 2025 (www.gov.uk) with a 12-month transition period (www.gov.uk). That verifies the individual's identity; it does not by itself verify the accuracy of ownership content. Open Ownership's verification principles recommend cross-checking against other authoritative systems and registers (openownership.org), which is another route to documented independent origin.

Keep records reusable

Store the ledger and graph with each diligence file so a later reviewer can see which origin supported which claim. Licensing terms for retaining provider metadata and excerpts vary by contract; counsel should review them, and this report gives no legal advice.

Implications and Future Directions

For buyers. Contracts can ask for origin metadata as a deliverable rather than a feature. A provider that supplies returned URL, publisher and document ID per field lets a buyer classify dispositions without a trial for every refresh. The Open Ownership standard shows what such metadata can look like: it separates the party asserting information from the publisher, with an assertion field for the people or organisations providing the information (standard.openownership.org), and a source type for information taken from an official register (standard.openownership.org).

For AI-assisted research. Agents that cite sources will often return aggregator or syndicated URLs. Because retrieval systems may point to syndicated versions rather than originals, as the Tow Center study documents (www.cjr.org), an agent's citation list should be read as a set of returned URLs to be mapped to origin groups, not a count of independent sources.

Open questions. Whether field-level origin metadata can be made standard across commercial providers, how often undisclosed upstream licences make independence unknowable, and what share of agreeing fields in real trials fall into each disposition are all unmeasured. A controlled multi-provider trial, run with lawfully obtained sample data and reviewed by the buyer's counsel, is the missing evidence.

Limitations of this report. No vendor export was inspected, no multi-provider trial was run, and no dependence or accuracy benchmark was produced. Vendor pages carry no as-of date, so statements reflect the pages as checked in November 2026.

Frequently Asked Questions (FAQs)

Are private equity data sources independent of each other?

Not by default. Providers describe overlapping source families such as regulatory filings, news and investor contributions, but overlap by type is not overlap by instance. Independence has to be documented per provider and field. Unknown is a legitimate result.

Are data providers using the same source?

Public pages cannot answer this for specific fields. PrivCo describes licensed datasets (see Table 1), and Crunchbase says third-party datasets are available on its platform (support.crunchbase.com); neither names the upstream for a given field. A trial with origin metadata is the way to find out.

How can a team test data provider independence?

Build a field-origin ledger, draw a claim-level origin graph, classify each pair into one of four dispositions, and compute agreement rate, documented-origin coverage, known origin groups and unknown-lineage share, each with denominators.

How should teams verify company data across providers?

Check the origin first, then the values. Where an authoritative record exists, go to it. The SEC's note that Form D data are derived from information provided by registrants (www.sec.gov) is a reminder that even official copies of self-reported data carry the reporter's limits.

What does data lineage mean for private equity research?

It means being able to trace a screen value back through publisher and document to an original observation, with dates, so that a reviewer can tell what supports a claim. The PROV model's derivation, quotation and revision relationships give a vocabulary for it.

How accurate is private company data?

Available evidence is domain-specific. The NETS analysis shows heavy imputation of small-firm employment (www.federalreserve.gov), and PrivCo describes labeling estimates (www.privco.com). No public benchmark measures ownership or mandate-fit accuracy across current commercial providers.

What belongs in a private equity research stack for deal sourcing?

A stack combines discovery platforms, authoritative records (filings, registries, entity identifiers) and an evidence layer that records origin. Dakota's guide notes that most private equity performance figures trace back to a handful of primary sources (www.dakota.com), a reminder that stack breadth and origin breadth differ.

Conclusion

Agreeing values can be a single evidence stream. The practical response is not to distrust every provider but to classify each provider and field pair as sharing a documented origin, having independently documented origins, partially sharing derivation, or being opaque. A second provider earns credit as corroboration only on fields where its origin is documented and separate, and it can still earn its place through access, normalization or freshness where it is not.

The method asks for modest evidence: a trial sample, a ledger and a graph. Its outputs are counts with denominators, lower bounds for known origin groups, and an acceptance gate for consequential claims in which unresolved remains an allowed answer. Undisclosed upstream licences mean independence will

sometimes be unknowable, and the report treats that as a finding to record. The next step is a controlled trial on authorized data, which would replace the synthetic numbers here with measured ones.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.w3.org/TR/prov-primer
2. https://www.vldb.org/pvldb/vol2/vldb09-335.pdf
3. https://api.opencorporates.com/documentation/API-Reference
4. https://www.privco.com/methodology
5. https://altss.com/methodology.md
6. https://preqin.com/data/our-data
7. https://arxiv.org/pdf/1503.00303
8. https://www.sec.gov/resources-small-businesses/capital-raising-building-blocks/form-d
9. https://www.opimas.com/research/917/detail
10. https://pitchbook.com/research-process-faqs
11. https://support.crunchbase.com/hc/en-us/articles/360009616013
12. https://www.sec.gov/dera/data/form-d
13. https://lunadong.com/publication/dependence_vldb.pdf
14. https://standard.openownership.org/en/0.4.0/standard/modelling/dates-guidance.html
15. https://www.gov.uk/guidance/people-with-significant-control-pscs
16. https://arxiv.org/pdf/1505.02463
17. https://pitchbook.com/help/pitchbooks-research-process
18. https://preqin.com/our-products/company-intelligence/methodology
19. https://www.sourcescrub.com/product/data
20. https://grata.com/private-company-data
21. https://www.nber.org/system/files/working_papers/w22500/w22500.pdf
22. https://www.nber.org/system/files/working_papers/w17874/w17874.pdf
23. https://www.w3.org/TR/prov-dm
24. https://prnewswire.com/pr-distribution-and-placement
25. https://developers.google.com/search/docs/crawling-indexing/canonicalization-troubleshooting
26. https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-1-data-who-is-who
27. https://ludovicphalippou.substack.com/p/navigating-private-equity-data-what
28. https://medium.com/@andreretterath/what-s-the-best-startup-vc-database-8237fc189830
29. https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/11/using-crunchbase-for-economic-and-managerial-research_9ba53344/6c418d60-en.pdf
30. https://www.federalreserve.gov/econres/feds/files/2019034pap.pdf
31. https://www.sec.gov/corpfin/form-d-filing

32. https://resources.companieshouse.gov.uk/disclaimer.shtml
33. https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom
34. https://www.gleif.org/en/about-lei/common-data-file-format/level-2-data-reporting-exceptions-format
35. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php
36. https://dealroom.net/blog/examples-of-ai-in-m-a
37. https://grata.com/resources/ai-in-mergers-and-acquisitions
38. https://www.niemanlab.org/2025/03/ai-search-engines-fail-to-produce-accurate-citations-in-over-60-of-tests-according-to-new-tow-center-study
39. https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/Guidance-Beneficial-Ownership-Legal-Persons.html
40. https://www.gov.uk/guidance/verify-your-identity-for-companies-house
41. https://openownership.org/principles/verification
42. https://standard.openownership.org/en/latest/standard/reference.html
43. https://www.dakota.com/resources/blog/where-to-find-private-equity-fund-performance-data

THE PUBLISHER

About Docket

Docket provides deal-origination research software and managed research for private equity firms. We help investment teams investigate acquisition targets using structured screening criteria, retained sources and reviewable company evidence. Teams can work through a self-serve platform or use managed research, depending on how they want research delivered.

Research against a defined mandate

A useful target list needs more than company names. Docket focuses on the questions that determine whether a company fits an investment mandate, including the evidence needed to support or qualify each answer. Our research approach makes the connection between screening criteria, source material and conclusions visible to the team reviewing the work.

Triage, Scout and Audit

Docket's three named research agents perform complementary tasks. Triage resolves company identity and screens fit. Scout collects sourced answers against the mandate. Audit checks retained evidence, addresses contradictions and leaves unsupported answers visibly unresolved. This structure helps reviewers distinguish established findings from missing information and questions requiring further investigation.

Evidence that supports investment-team judgment

Our research library covers market mapping, screening criteria, private-company data, succession and ownership, source evaluation and evidence standards. These resources explain the methods and limitations behind origination research. Findings support a team's judgment; they do not establish that a company is for sale or guarantee a transaction or investment outcome.

Work with Docket

Visit Docket to explore the platform, managed research and the current contact path. Read about screening criteria, evidence standards, AI research agents and market mapping.

Public examples are illustrative unless explicitly identified otherwise. Research preparation and authorized outreach are separate activities; confidential target lists and customer outcomes should never be inferred from an educational example.

Website: https://docket.capital

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.