Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

How to Evaluate Private Company Data Vendors: A Checklist

> 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 · 2026-10-29 · how to evaluate private company data providers · private company data accuracy · data vendor due diligence · private company data coverage

Original article: https://docket.capital/articles/vendor-evaluation-questions-private-company-data



In this article

1. What this checklist is for
2. Define coverage before comparing vendors
3. Sourcing: where does each field come from?
4. Verification and update cadence
5. Read the published numbers critically
6. Entity resolution and ownership data
7. Licensing, export and redistribution terms
8. Designing a sample test
9. A scoring worksheet and decision record

What this checklist is for

A data subscription is often bought on a demo, a record count and a price. None of the three shows whether the records a mandate needs exist, sit on the right legal entity, and were checked recently enough to act on.

This article gives 14 vendor-neutral questions and a sample-test design. It does not rank providers; for a list, see best private company data providers for PE. PitchBook's and Crunchbase's published method statements serve as worked examples; everything quoted is the vendor's self-description as published on 2026-10-28, not an independent finding.

Published descriptions tell you what to ask; only a test against records you already know tells you what you are buying.

Define coverage before comparing vendors

Before any demo, write down what "coverage" means for the mandate in four dimensions:

Dimension	Write down	Example of a precise definition
Entities	Which kinds of organization count	Operating companies with a US headquarters; exclude holding shells and branch locations
Geography	Where the entity must be located or operate	Headquartered in the 12 states of the thesis
Size	Which size band, measured how	20-250 employees, any source, with the source named
Fields	Which attributes must be populated and usable	Legal name, website, ownership type, founder or CEO name, employee estimate, last-verified date

Two numbers then matter, and they rarely agree: the entity count (records matching the filters) and the usable record count (records that are the right entity, in scope, with acceptable values in the required fields).

Duplicate records widen the gap. A Census Bureau overview of record linkage notes that duplicates in a list "can have the effect of erroneously inflating estimates of the number of entities in different categories," and gives business lists by industry as the example (Winkler, Census Research Report 2006-2). Counts that include duplicates, branches and dormant entities overstate the workable universe. To estimate how much of a target universe a source captures, compare it with an independently built list; the method is set out in when a target universe is complete.

Questions to ask

1. For our exact filter set, how many records do you return, and how many of those are distinct legal entities rather than locations, brands or duplicates?
2. What are your inclusion and exclusion rules for a company to receive a profile?

Question 2 has concrete answers in at least one published case. Crunchbase's accuracy page states that, to qualify, a company needs to demonstrate "structured entities," "scalable models" and "ecosystem impact," and that "hyperlocal storefronts, walk-in service businesses, and individual point-of-sale setups" are explicitly out of scope (Crunchbase, "How Accurate Is Crunchbase Data?"). Read inclusion rules against the thesis: owner-operated service businesses a buy-and-build targets may fall outside a vendor's stated scope by design. That is a fit question, not a defect.

Sourcing: where does each field come from?

Vendors describe sourcing for the whole dataset; buyers need it per field, because a funding round, an employee count and an owner's name are collected in different ways and age at different rates.

Two published self-descriptions

PitchBook's research-process page states: "Our research is based on public and self-reported data." It describes a sequence: web crawlers "capture information from secondary sources such as news articles, regulatory filings, websites, and press releases"; "Natural language processing and machine learning technology organizes the data and filters out irrelevant information"; "specialized data teams collect, calculate, and confirm key information"; a "Quality Assurance team uses preventative validation and manual reviews"; and "The Primary Data Operations team communicates directly with a firm's management and gathers additional key details, on a quarterly basis" (PitchBook, Research Process).

Crunchbase's support page states: "Rather than relying on web searching or scraping the web, the vast majority of our data comes directly from our venture partners and our active community of contributors." It cites "4,000+ members" in its Venture Program, who "submit monthly portfolio updates," "600,000+ executives, entrepreneurs, and investors" who update profiles, and "Over 400+ algorithms" that "search for and validate our data on a daily basis" (Crunchbase Support, "Where does Crunchbase get their data?"). The accuracy page lists five inputs: proprietary ingestion systems, external sources including "more than 1,000 news outlets and government filings," analyst validation, user engagement signals, and market experts (Crunchbase accuracy page).

Each named mechanism becomes a field-level question. "Self-reported" means some values came from the company itself; which fields, and are they flagged? "Portfolio updates" from investors suggest strength where investors are present; what is the source for companies with no institutional investor, which describes much of the founder-owned lower middle market? That is an inference from the stated mechanism, not a claim about either vendor's coverage; test it.

Questions to ask

3. For each field we require, what is the primary source type (filing, company website, self-report, investor submission, model estimate)?
4. Is the source type visible per record, or only described in aggregate?
5. Which fields are modeled or estimated (for example employee counts or revenue), and how is an estimate distinguished from a reported value in the export?

Question 5 matters most for size fields: a reasonable modeled revenue figure can still be the wrong input for a hard threshold. Ask for the estimate and the reported value in separate columns, so a screen can treat them differently.

Verification and update cadence

Crunchbase's accuracy page describes a "Multi-Point Independent Verification Standard": "Before a company profile is created or updated, the change has to be checked against at least two independent, official sources." Accepted sources listed include the company's official website and verified address, "A corporate registry or government database," "A verified corporate social profile or employer listing," and "Credible news coverage, M&A activity, or funding events." The page adds that the standard "applies to individual data points, not just full profiles" (Crunchbase accuracy page). PitchBook's page describes quarterly direct contact by its Primary Data Operations team with "a firm's management" (PitchBook, Research Process).

Neither statement tells you when a given field on a given record was last confirmed, which is what a deal team needs before calling an owner.

Questions to ask

6. Does each field carry its own last-verified date in the platform and in the export, or only a profile-level "last updated" date?
7. When a verification rule requires more than one source, can we see which sources confirmed a given value?
8. What share of records in our filter set had a required field confirmed in the last 12 months, and how is "confirmed" defined?

If the vendor cannot answer question 8 for your filter set, measure it yourself in the sample test below.

Read the published numbers critically

Vendor figures can drift between pages, or within one page, without intent to mislead. Ask for the definition and date behind every number used in the decision.

A worked example, as published on 2026-10-28. Crunchbase's accuracy page (marked "Last Updated: October 20, 2026") contains two statements about contributor activity:

Location on the page	Statement as published
"Where Does Crunchbase Get Its Data?" section	"Contributor activity accounts for roughly 14.2% of all data updates processed across the platform, a small share of the total."
FAQ, "Does Crunchbase Use Contributor-Submitted Data?"	"Contributor activity accounts for roughly 1.6% of all data updates processed across the platform."

The same page refers to "a network of 700,000+ active contributors" (Crunchbase accuracy page), while the support page refers to "600,000+ executives, entrepreneurs, and investors" who update profiles (Crunchbase Support).

Different periods, definitions or an unrevised passage could explain this; which applies is unknown, and nothing here bears on data accuracy. The point is general: any vendor's headline figures (record counts, update volumes, source shares) need a definition and a measurement date before they enter a memo or a procurement scorecard.

Question to ask

9. For each figure in your materials that we are relying on, what is the definition, the measurement period and the date it was last updated?

Entity resolution and ownership data

For PE origination, the expensive errors are not a wrong website or a stale headcount. They are records attached to the wrong legal entity, two records for one company counted as two targets, and a subsidiary presented as an independent founder-owned business. All three are entity resolution problems.

Winkler's Census overview describes the record linkage model "introduced by Newcombe (1959, 1962) and given a formal mathematical foundation by Fellegi and Sunter (1969)," and notes that "Because much of the data is based on textual information such as names and addresses," much of the progress has come from methods such as "string comparison for accounting for typographical error" and parsing names and addresses into comparable components (Winkler 2006). A Census Center for Economic Studies paper on linking survey respondents to employers in the Census Business Register describes linkage "in the absence of unique identifiers" as challenging, uses "a supervised machine learning model to probabilistically link" records, and uses multiple imputation "to propagate uncertainty from the linkage step into subsequent analyses" (Abowd et al., CES-WP-19-08, 2019).

The lesson for a buyer: company matching is probabilistic, and a process can report how uncertain a match is. Ask whether the vendor's does.

Questions to ask

10. How is a record matched to a legal entity (registry identifier, name and address matching, manual review), and is the matched identifier included in the export?

11. How are parent and subsidiary links built and dated, and can we export the full corporate family for a record?
12. When two records may describe the same company, how is that resolved, and can we see match confidence or a review flag?

Check ownership type separately: "independent" may mean only "no known institutional investor in our data," which is weaker than "verified as owner-operated."

Licensing, export and redistribution terms

Questions only; nothing here describes any vendor's contract. Read the agreement and involve counsel where answers affect use.

Topic	Question for the vendor (confirm in the contract)
Export limits	How many records and fields can be exported per period, per seat or per account?
Internal use	Who counts as a permitted user: deal team only, operating partners, portfolio companies, outside advisors?
Derivative works	If we enrich exported records with our own research, who owns the combined record?
CRM sync	May exported data be stored in our CRM, and does that right survive termination?
AI and model use	May the data be used in internal AI tools, prompts or model training, and under what limits?
Redistribution	May excerpts appear in investment committee memos, lender materials or LP reports?
Retention at termination	What must be deleted, by when, and does that include records already merged with our own notes?
Audit rights	Can the vendor audit our usage, and what records must we keep to show compliance?

Questions to ask

13. Which of the items above are governed by the master agreement, which by an order form, and which by online terms that can change during the term?

If the firm expects to keep research notes built on vendor records after a non-renewal, confirm in writing that it may.

Designing a sample test

Build the test set

- Pull a test set of targets the team already knows well: companies met, diligenced, or researched from primary sources. A set of 100 to 200 records is a workable starting point for a first comparison; the right number depends on the tolerance you need. Sampling arithmetic and stratification are covered in how many records to sample for an audit.

- Include known hard cases on purpose: companies with a trade name different from the legal name, multi-state registrations, recent ownership changes, and subsidiaries that look independent.
- Add a small set of companies you know are out of scope, to test whether filters exclude them.
- Record your own "answer key" for each required field, with the primary source and the date you confirmed it. Fields without a confirmed answer stay marked "open" and are excluded from scoring.

Score each field

For each record and field, classify the vendor value against the answer key:

- **Match:** same value, allowing a written tolerance (for example employee band rather than exact count).
- **Mismatch:** a value is present and conflicts with the answer key.
- **Missing:** no value.
- **Wrong entity:** the record describes a different company than intended.

Track wrong-entity results separately: one wrong-entity record contaminates every field on it.

Illustrative scorecard (fictional)

Invented to show the layout; "Vendor A" and "Vendor B" are not real providers.

Field (150-record fictional test set)	Vendor A match / mismatch / missing	Vendor B match / mismatch / missing	Notes
Legal entity correctly identified	140 / 10 / 0	130 / 5 / 15	B returned no record for 15 trade-name cases
Website	145 / 5 / 0	135 / 0 / 15	
Ownership type	100 / 30 / 20	90 / 20 / 40	A labels 30 subsidiaries "independent"
Founder or CEO name	110 / 20 / 20	80 / 10 / 60	
Employee band	90 / 50 / 10	70 / 40 / 40	
Field-level last-verified date shown	No	Yes	

In this fictional case, neither vendor "wins" on the table alone. Vendor A fills more fields; Vendor B leaves more gaps but makes fewer ownership errors. A visibly open gap can be researched, while a confident wrong value may never be questioned. That trade-off is the deal team's judgment, recorded in the decision record below.

Question to ask

14. Will you run our test set, or give us trial access that permits it, under the same terms and export format we would have as a subscriber?

A scoring worksheet and decision record

A scorecard serves one mandate, not a verdict on a vendor. Fix the weights before the test runs.

Criterion	Weight (set before testing)	Evidence	Score (1-5)	Reviewer
Usable record count for our filters (Q1-Q2)		Test exports, overlap check		
Field-level source visibility (Q3-Q5)		Demo screenshots, export columns		
Field-level verification dates (Q6-Q8)		Export columns, sample results		
Definitions and dates for cited figures (Q9)		Vendor written answers		
Entity and ownership accuracy (Q10-Q12)		Wrong-entity and ownership results		
Licensing fit (Q13)		Contract review notes		
Test access on subscriber terms (Q14)		Trial terms		

Then write a one-page decision record:

- **Mandate and scope:** the coverage definition from the first section.
- **Test set:** size, how it was drawn, hard cases included, date of the answer key.
- **Results:** the scorecard, with wrong-entity results stated separately.
- **Vendor answers:** written responses to the 14 questions, dated.
- **Open items:** anything unanswered, marked open rather than assumed.
- **Decision and owner:** the choice, the person signing off, and the date to re-test (for example at renewal or when the investment criteria change).

Filing the vendor's answers with the test results lets the next renewal start from evidence.

Docket works on the other side of this problem: its Triage, Scout and Audit agents research a client's own target list and keep each answer tied to a quoted, retained source page, with unresolved fields left visibly open.

Source links

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

1. https://www.census.gov/content/dam/Census/library/working-papers/2006/adrm/rrs2006-02.pdf
2. https://about.crunchbase.com/blog/crunchbase-data-accuracy-and-verification
3. https://pitchbook.com/research-process
4. https://support.crunchbase.com/hc/en-us/articles/360009616013-Where-does-Crunchbase-get-their-data
5. https://www2.census.gov/ces/wp/2019/CES-WP-19-08.pdf

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.