Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

How to Know When a Private Equity Target Universe Is Complete

> 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 · 2026-09-26 · private equity target universe · target universe coverage · market mapping · deal sourcing · capture recapture · overlap analysis · entity resolution · source yield · stopping rule

Original article: https://docket.capital/articles/pe-target-universe-completeness-overlap-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Implementation Considerations and Process Changes
5. Overlap Estimation and the Stopping Rule
6. Data Analysis and Evidence
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

A private equity target universe is **operationally complete** when its definition is frozen, every accepted candidate has a resolved identity and source history, the next plausible source has an unattractive expected yield for the mandate, and the remaining uncertainty is stated explicitly. That is a decision standard, not a claim that every eligible firm has been found. A statistical firm count, a mandatory register, and an overlap estimate are different kinds of evidence. U.S. Statistics of U.S. Businesses (SUSB) reports firms and establishments, while County Business Patterns (CBP) reports establishments; neither is automatically a list of acquirable companies. (www.census.gov) If no matching denominator exists, the observed union is a verified **minimum**, and any estimated total should be a sensitivity scenario.

The proposed operating test has **four parts**. First, define the acquisition unit, geography, size range, capability rule, and observation window. Second, preserve a source-membership matrix after identity resolution; false and missed links change the overlap on which an estimate depends. (www.ons.gov.uk) (www150.statcan.gc.ca) Third, measure each source's accepted incremental targets against review time and purchase cost. Fourth, use multiple source pairs and segments to show a range rather than a single coverage percentage. The World Health Organization (WHO) lists closure, independence, homogeneous inclusion, and accurate matching among the conditions for a valid two-list estimate; it warns that two-sample methods are often unsuitable where lists depend on one another. (applications.emro.who.int)

In the **fictional worksheet** below, three lists contain **100**, **90**, and **70** in-scope entities. Their deduplicated union is **180**. Pairwise Chapman estimates imply scenario totals of about **223 to 248**, so observed coverage spans about **73% to 81%** when each pair's estimate is compared with the full union. These are calculations from invented data, not a measured PE market. The source pairs overlap differently, and no two-list calculation can identify a segment whose firms have essentially no chance of entering any list. Methodological work finds estimates become unstable when capture probabilities concentrate near zero. (link.springer.com) (arxiv.org)

The practical stopping decision is to compare the next source's likely accepted additions and review burden with the mandate's cost of omission, then report a **coverage range, source scope, unsearched segments, and date**. A low marginal yield alone is insufficient if several sources share the same bias. A two-list interval measures sampling uncertainty under model assumptions; it does not absorb dependence, poor matching, or an undefined market boundary. The Office for National Statistics (ONS) builds a separate coverage survey and

still recognizes dependence bias, which shows the care required before applying the method to ordinary company lists. (www.ons.gov.uk) Commit consequential population estimates to statistical review; where assumptions cannot be defended, report the observed lower bound and the unknown remainder. (arxiv.org)

Introduction and Background

The question “is the target universe complete?” often appears after a team has already assembled a long spreadsheet. It cannot be answered by the length of the list. It asks whether the list represents the **defined acquisition opportunity**, whether more searching will find materially different candidates, and how much uncertainty remains. A fragmented niche without a compulsory, exhaustive register is especially hard because there is no directly observable count of eligible firms. Docket's existing market-mapping guide describes statistical counts and source unioning; this report addresses the next decision when those routes do not yield a defensible denominator.

The first distinction is between a **population**, a **frame**, and an **observed union**. The population is all entities meeting the mandate on an agreed date. A frame is a source that could enumerate some of them, such as a license file, association directory, exhibitor list, or commercial dataset. The union is the set that survives matching and eligibility review. Statistical business registers often use enterprises as a unit, but a target mandate may call for an operating group, a legal entity, or a carve-out. Eurostat describes an enterprise as potentially comprising multiple legal units, so even an official enterprise count requires a unit crosswalk. (ec.europa.eu)

In U.S. work, the North American Industry Classification System (NAICS) classifies **establishments** for statistical reporting. (www.census.gov) CBP counts establishments, which need not equal companies; SUSB publishes both firm and establishment counts; the Quarterly Census of Employment and Wages (QCEW) measures employment and wages associated with unemployment-insurance coverage. (www.census.gov) (www.bls.gov) QCEW's published coverage of more than **95% of U.S. jobs** is not a claim that it names more than 95% of acquisition targets. (www.bls.gov) A niche defined by a capability that spans codes, ownership structures, and self-employed operators may not align with any of these aggregate counts. This mismatch is the reason to keep denominator evidence separate from list-discovery evidence.

Coverage should therefore be phrased as a **conditional answer**: “Of the companies meeting this rule that our named sources could reach as of this date, we have accepted this many; under these overlap assumptions, the estimated total is this range; these segments remain unmeasured.” The point of the method is to make the next research or dataset purchase decision auditable, not to certify an unknowable total.

Key Changes

Define a target unit before counting

A reproducible universe begins with one row per acquisition unit. Specify whether the row represents an independent operating business, a controlled group, or a legal entity. Record geography, minimum and maximum size, ownership exclusions, capability evidence, and the **as-of window**. The same brand can cover

several legal entities; several locations can belong to one enterprise. The Census Bureau explicitly notes that an establishment is not necessarily identical with a company. (www.census.gov) A source can be valuable even when it counts a different unit, but its count cannot silently become the denominator.

Use a written **capability rule** rather than a keyword or code alone. For example, a fictional mandate might require independent firms that maintain specialized production equipment for third-party manufacturers in three states, within a specified employee band, on **September 26, 2026**. The rule should say whether resellers, sites owned by parent groups, and mixed-service firms qualify. Any later rule change needs a version and a re-review queue. Industry-code misclassification can create undercoverage and overcoverage in subsets, according to United Nations statistical business-register guidance. (www.census.gov) (unstats.un.org)

Before importing a source, freeze these scope fields in the mandate worksheet:

- **Unit:** State whether the target is an operating company, group, or legal entity.
- **Geography:** List included jurisdictions and the rule for multi-state operators.
- **Size:** Define the metric, band, reference period, and treatment of missing values.
- **Capability:** State what evidence qualifies and what merely triggers review.
- **Window:** Fix the observation date and the acceptable age of source records.

Treat identity resolution as part of measurement

Normalize names, addresses, domains, and official identifiers, then review ambiguous pairs. Exact matching, probabilistic matching, and clerical review are recognized linkage approaches; ONS warns that a raw match rate says little about linkage quality. (www.ons.gov.uk) Over-aggressive normalization can increase false positive matches, while missed links split one firm into two source histories. (www.ons.gov.uk) (www.gov.uk) Keep the original source row, normalized fields, chosen entity ID, merge rationale, reviewer, and correction history. Where a registry number exists, retain its issuing jurisdiction and original formatting; ONS notes that even leading zeros can determine whether references match. (untp.unece.org) (www.ons.gov.uk)

For overlap analysis, the effect of a matching error is directional. A **false merge** raises apparent overlap and can drive a two-list estimated total downward. A **missed match** lowers overlap and can drive it upward. This follows directly from the denominator of the Chapman formula below, and Statistics Canada identifies linkage error as a source of bias in capture-recapture estimates. (www150.statcan.gc.ca) (link.springer.com) Use a target audit sampling worksheet to review separate, representative samples of accepted links and unmatched records or candidate nonlinks. Check accepted links for false links (precision), search unmatched records for missed true links (recall), and document each sampling frame, sample size, and extrapolation rather than relying on a high automated match rate. (www.ons.gov.uk)

The identity review log should retain:

- **Original row:** Preserve the name, address, site, identifier, and source timestamp.
- **Canonical ID:** Assign one key to the chosen acquisition unit.
- **Candidate links:** Keep potential matches and the evidence tested.
- **Decision:** Record merge, split, or unresolved status with a reviewer.
- **Audit sample:** Measure false and missed links in reviewed examples.

Replace a single completeness claim with an evidence ladder

Table 1 distinguishes the denominator evidence a team has from the workflow it uses to build the candidate union. It includes Docket as a direct provider of target research, with the scope its own site describes; a research service does not itself supply a statistical population count.

Route	What it can establish	Decision use and limit
Statistical firm denominator	A count for a defined official unit, geography, code, size band, and reference year. (www.census.gov)	Compare an eligible union only after reconciling unit and scope. SUSB release timing can lag its reference year by about **2 to 2.5 years**. (www.census.gov)
Mandatory register	Named entities covered by a specified licensing or filing rule. (www.sba.gov)	Audit legal scope, exemptions, and date; a register can miss activities outside that rule. U.K. licensing guidance cautions that an inventory may omit licenses. (www.gov.uk)
Overlap estimate only	A scenario total inferred from matched, in-scope list memberships. (applications.emro.who.int)	Report pair and stratum sensitivity; the total is unidentified without assumptions about unseen firms. (arxiv.org)
Docket target research	Docket says teams may import a list or request one built around their thesis, with identity checks and sourced answers.	A direct research option for building and auditing the numerator; it does not turn an overlap scenario into a census.

The decision path is **not additive**. A good register may make a two-list estimate unnecessary. A statistical count can provide an external plausibility bound without being an exact denominator for a capability-defined market. Docket's own description of source excerpts and collection dates supports record-level auditability, while the population claim still depends on the market definition and source coverage.

Implementation Considerations and Process Changes

Build a source-membership matrix

Create one canonical row per accepted entity and a binary column per searched source. Retain rows rejected as duplicates or out of scope in a separate review log, since those rows explain raw-to-accepted conversion. A minimal matrix contains `entity_id`, `unit_type`, `jurisdiction`, `size_band`, `capability_status`, `source_A`, `source_B`, `source_C`, `first_seen`, `last_verified`, and `identity_review`. A multiple-list method treats the observed combinations as an incomplete contingency table; with three lists, **seven** nonempty patterns are observable and the all-absent cell is hidden. (pmc.ncbi.nlm.nih.gov)

For each source-membership flag:

- **Value:** Store one for a confirmed appearance and zero otherwise.
- **Eligibility:** Keep source appearance separate from in-scope acceptance.
- **Provenance:** Link the membership to the original row or page.
- **Timing:** Retain collection and verification dates.
- **Corrections:** Recompute patterns after a split or merge.

Preserve provenance at the field level. A source membership flag answers “where was this firm found?”, while a capability citation answers “why is it eligible?”. Those are different questions. A commercial data row that names a firm but lacks current evidence of the required capability should remain a candidate until reviewed. Research on the Orbis commercial database highlights its use for analyzing top performers and multinationals, illustrating why one vendor file should not be presumed representative of every private-company segment. (www.jonathantimmis.com)

Measure incremental yield in source order

For each source, record **raw rows, unique resolved entities, in-scope entities, new accepted entities after union, review minutes, and license cost**. “New” is order-dependent: if an association directory is processed before an exhibitor list, the association gets credit for shared firms. Recompute the curve for at least one different source order, or attribute shared names separately. The curve plots cumulative accepted unique firms against cumulative effort or money. Such accumulation plots are established in inventory research, but the plateau is an operational signal, not proof that the unobserved cell is empty. (besjournals.onlinelibrary.wiley.com) (www.cochrane.org)

The yield sheet should calculate:

- **Duplicate rate:** Raw rows minus unique resolved entities, divided by raw rows.
- **Eligibility rate:** In-scope entities divided by unique resolved entities.
- **Marginal yield:** New accepted entities after union for that source.
- **Review effort:** Minutes spent screening and resolving identity.
- **Marginal cost:** Source fee plus labor, divided by accepted additions.

Table 2 is a **fictional example**. Labor is assumed at **$75 per hour** and source fees are invented. Each row's incremental cost includes its labor and fee; no PE market benchmark is implied.

Fictional source, in order	Raw rows	Unique entities	In scope	New accepted	Review minutes	Source fee	Cost per new accepted
A, trade directory	125	112	100	100	720	$0	$9.00
B, event exhibitor list	118	105	90	50	600	$1,000	$35.00
C, specialist dataset	94	80	70	30	450	$600	$38.75

The union grows **100 → 150 → 180** while the accepted increment falls **100 → 50 → 30**. That does not establish saturation: the three sources may all favor visible, larger businesses. The marginal cost calculation is `(source fee + review minutes / 60 × hourly rate) / new accepted`. If a proposed fourth dataset costs $2,000 plus 8 analyst hours and a blinded trial yields 20 accepted new firms, the fictional expected cost is **$130 per accepted addition**. The decision depends on whether those additions matter to the mandate and whether a materially different segment is uncovered.

Inspect overlap before fitting a model

Show pairwise intersections and the full source-membership patterns. High overlap can mean broad coverage, but it can also mean one list copied or drew from another. Low overlap can mean complementarity, or it can result from different geographies, size bands, stale records, and missed identity links. WHO's independence and homogeneous-capture assumptions are exactly why these interpretations must be tested before estimating a total. (applications.emro.who.int)

Where sources refer to different eligible populations, harmonize them or exclude non-comparable records. A Centers for Disease Control and Prevention (CDC) example excluded cases that could not have appeared in both systems before calculating overlap. (www.cdc.gov) A target list covering only one state cannot be paired uncritically with a nationwide list to infer a national total. Source scope, acquisition unit, and observation window must match.

The assumption checklist should record a verdict and evidence for each item:

- **Closure:** Did the population remain reasonably stable within the window?
- **Independence:** Could inclusion on one list affect another list?
- **Inclusion:** Do size and geography change a firm's chance of appearing?
- **Matching:** Were ambiguous identities reviewed and errors sampled?
- **Case rule:** Was the same eligibility definition applied to all lists?

Overlap Estimation and the Stopping Rule

Use Chapman as a sensitivity calculation

For two comparable lists, let `n1` and `n2` be their eligible unique counts and `m` their matched overlap. The **Chapman estimate** is $\hat{N}$ `= ((n1 + 1) × (n2 + 1) / (m + 1)) - 1`. The observed two-list union is `U = n1 + n2 - m`; the estimated unseen number is $\hat{N}$ `- U`; observed-to-estimated coverage is `U /` $\hat{N}$. Chapman is a small-sample correction to the elementary ratio, with published variance and interval methods. (link.springer.com) (online.stat.psu.edu) (search.r-project.org) These expressions are arithmetic on the included records; they do not repair biased source selection.

For a fixed pair of list sizes, a smaller `m` produces a larger estimated unseen cell. That relationship is useful diagnostically: it makes the model's dependence on accurate matching visible. (hrdag.org) A confidence interval based on Chapman variance can be reported **conditional on closure, comparable capture probabilities, independent sources, and correct links**. Penn State presents an approximate confidence interval, while a methods paper notes that independence cannot be tested using only two lists. (online.stat.psu.edu) (pmc.ncbi.nlm.nih.gov) Thus an apparently narrow interval is no defense against a shared directory bias.

Stratify, then compare source pairs

Divide the universe into segments with plausibly different discovery probabilities, for example geography and company size. Estimate within a stratum only where the source windows and target rule match. ONS stratified its Census Coverage Survey using geography and a hard-to-count measure; United Nations guidance groups

units with similar capture probabilities. (www.ons.gov.uk) (unstats.un.org) Very fine slices can leave too few overlaps for stable estimation, so retain sparse strata as “unknown” or combine them based on a defensible rule. (pmc.ncbi.nlm.nih.gov) (unstats.un.org)

Run each credible source pair within each viable stratum. Large divergence is a warning to audit source dependence, capture heterogeneity, and matching. Three or more sources allow log-linear models with interactions, but model selection and the unseen cell still depend on assumptions. (pmc.ncbi.nlm.nih.gov) (arxiv.org) For an investment-committee claim or a sizable dataset purchase, ask a statistician to review the design, not merely the spreadsheet formula. If no identifying assumption is defensible, a methodological paper recommends making no population-size estimate from that dataset. (arxiv.org)

Stop when the next search is unlikely to change the decision

A stopping rule needs **four recorded inputs**: the recent accepted marginal yield, the all-in review cost, the range of plausible unseen firms, and the mandate's omission risk. Set the threshold before evaluating the next source, then pilot that source on a comparable, blinded slice. The threshold is a management choice, not a universal PE statistic. Search-method guidance outside PE notes limited formal evaluation of stopping approaches, and government data-quality guidance says measures should be tied to the intended use and its risk. (www.cochrane.org) (www.gov.uk)

Proceed if the source reaches a neglected segment, changes the top target set, or yields accepted additions at an acceptable cost. Pause if repeated independent sources add little, the high-risk strata have been reviewed, and further search is unlikely to affect the acquisition thesis. Escalate if the sources share a discovery channel or the estimated range remains decision-changing. Always log what was searched, by whom, when, and with which terms; reproducibility matters as much as the headline count. (www.cochrane.org)

Document the disposition of each next-source proposal:

- **Buy or search:** Expected additions justify cost or reach a neglected stratum.
- **Pilot:** The source mechanism is promising but overlap is uncertain.
- **Audit:** Identity or case-definition errors could change the estimate.
- **Pause:** Marginal yield is low and residual risk is within mandate tolerance.
- **Reopen:** A thesis change, new register, or refresh trigger alters coverage.

Data Analysis and Evidence

The **fictional three-source matrix** below is deliberately small enough to audit. A, B, and C are different source lists; `1` means the same eligible entity appears in that source. These seven observed cells sum to **180**. The eighth cell, `000`, is not observed. (pmc.ncbi.nlm.nih.gov)

```
pattern       100   010   001   110   101   011   111   000
entities       45    40    30    25    15    10    15     ?
```

From this matrix, **A = 100**, **B = 90**, and **C = 70**. The pairwise intersections are **A∩B = 40**, **A∩C = 30**, and **B∩C = 25**. A Venn or UpSet chart can be built directly from the seven pattern counts. The incremental-yield curve from Table 2 is `100, 150, 180` after A, then B, then C. All figures in this paragraph are calculations from the fictional cells.

Table 3 applies the Chapman formula separately to each pair. The final coverage column compares the **same full observed union of 180** with each pair's estimated total; it is a scenario comparison, not a three-list estimator. (link.springer.com)

Fictional pair	List sizes	Overlap	Pair union	Chapman total	Total minus full union	Full union / total
A and B	100, 90	40	150	223.2	43.2	80.7%
A and C	100, 70	30	140	230.3	50.3	78.1%
B and C	90, 70	25	135	247.5	67.5	72.7%

The spread is **223.2 to 247.5**, with an implied remainder after the full union of **43.2 to 67.5**. It is a narrow illustration of pair sensitivity, not a defensible market bound. Positive list dependence could make the estimates too low; heterogeneous capture can move estimates in less predictable ways. ONS acknowledges dependence bias even in a deliberately designed coverage survey, while peer-reviewed work shows that near-zero capture probabilities can produce very high estimation risk. (www.ons.gov.uk) (arxiv.org)

The worksheet should have **six tabs or equivalent sections**: `Mandate` for unit and scope; `Source_rows` for original records and review outcomes; `Entity_matrix` for one canonical row and membership flags; `Patterns_and_yield` for the seven cells, overlap, and cumulative increments; `Estimates` for Chapman pairs, a conditional interval, and stratum scenarios; and `Decision_log` for assumptions, marginal cost, risk, and stop or continue decision. This is a copyable workbook specification, not a claim of measured market coverage. For each pair, compute the Chapman variance as `((n1+1)(n2+1)(n1-m)(n2-m))/((m+1)^2(m+2))`, then show an approximate interval based on its standard error **only if** its assumptions are judged credible. (search.r-project.org) (online.stat.psu.edu) For the fictional A-B pair, the variance is **390.5**, the standard error is **19.8**, and an illustrative normal approximation using `estimate ± 1.96 × standard error` gives about **184.4 to 261.9**. This is a conditional interval, not a bound on source bias.

For stratified scenarios, enter one row per geography and size band. Store each pair's `n1`, `n2`, `m`, estimated total, observed union, and scope flags. Sum scenario totals only across **mutually exclusive** strata, retaining `unknown` where overlap is too sparse. Do not hide a model failure behind a zero or a blank cell. The interval addresses random variation under the chosen model; a separate sensitivity column must show alternative pair, matching, and dependence assumptions. (pmc.ncbi.nlm.nih.gov) (arxiv.org)

Implications and Future Directions

The strongest committee language is a **bounded evidence statement**. It should disclose the mandate version, as-of date, observed unique eligible firms, sources and their scopes, review quality, recent incremental yield, and whether the total is a census-like count, register tally, scenario estimate, or unknown. Distinguishing those categories prevents a precise percentage from implying more knowledge than the

process has earned. Statistical agencies themselves manage unit definitions and frame quality: Eurostat describes enterprises as combinations of legal units, and U.N. business-register work records variation in units and inputs across countries. (ec.europa.eu) (unstats.un.org)

Future refreshes should preserve the same source-membership history. A newly discovered company should be tagged as a new entrant, a newly found old firm, or an identity correction. Otherwise the team cannot tell market change from discovery change. Keep a separate backlog of unresolved identity pairs and capability decisions; ONS says linkage errors affect coverage estimates, and U.K. data-quality guidance asks data custodians to communicate nonunique records. (www.ons.gov.uk) (www.gov.uk)

The next source purchase should be tested on **incremental accepted entities in under-covered strata**, not raw vendor row count. Commercial datasets can improve discovery while remaining poor denominators for particular firm segments; research on Orbis shows why representativeness must be tested for the exact use. (www.jonathantimmis.com) Where a mandatory register appears later, use it as an external check, but verify its jurisdiction, activity, exemptions, and observation date. U.S. licensing requirements vary by business activity and location; a general licensing finder may not enumerate every applicable license. (www.sba.gov) (www.gov.uk)

The method remains an aid to decisions under uncertainty. Three-list modeling, external bounds, and a statistician's review can improve a consequential estimate. They cannot make a zero-visibility segment appear in the data without a new discovery channel. Report the observed union as a lower bound and keep the unobservable remainder explicit. (arxiv.org)

Frequently Asked Questions (FAQs)

What is private equity target universe coverage?

It is the share of a **defined eligible population** represented by resolved, reviewed target entities. If an appropriate denominator is known, calculate accepted entities divided by that denominator after reconciling units and dates. If it is unknown, describe the observed union and conditional model scenarios; do not treat a commercial list size as the population. Official datasets distinguish establishments from firms, and some official microdata are restricted to qualified researchers rather than available as a public target list. (www.census.gov)

How should a PE team build a target universe?

Write the unit and capability rule, gather sources with different discovery mechanisms, preserve every original row, resolve identities, apply the same eligibility rule, and retain source membership. Then review source yield and overlap by segment. An acquisition target list is a screened output of that process, not a copy of a statistical industry code. NAICS classifies establishments, and statistical-business-register guidance warns that industry misclassification can distort subsets. (www.census.gov) (unstats.un.org)

Can capture-recapture prove the target list is complete?

No. It estimates an unseen cell under assumptions about closure, independence, comparable inclusion chances, and matching. WHO warns that two-source estimates often fail when lists are dependent; a two-list dataset cannot test independence internally. (applications.emro.who.int) (pmc.ncbi.nlm.nih.gov) The estimate

is best treated as a sensitivity case alongside the observed minimum, external bounds, and other source pairs. If important firms have near-zero probability of appearing in all chosen lists, the estimate can be unreliable. (arxiv.org)

How should overlap analysis affect the next dataset purchase?

Run a blinded sample of the proposed dataset through the same identity and eligibility process. Count **accepted new entities**, especially in the highest-risk strata, and divide source plus review cost by that count. Compare its source mechanism with existing lists: high overlap may reflect duplication, while low overlap may reflect genuine complementarity or matching problems. Government linkage guidance notes the trade-off between missed and false links, so resolve those before pricing the increment. (www.gov.uk)

When should searching stop?

Stop at a documented decision point when the next plausible source's expected accepted yield and strategic value no longer justify its cost, **provided** high-risk segments have been probed and the unknown remainder is disclosed. There is no validated universal percentage or cost cutoff for PE market mapping. Cochrane's broader search guidance notes limited formal evaluation of stopping methods, and use-specific data-quality guidance supports a mandate-specific threshold. (www.cochrane.org) (www.gov.uk) Reopen the search when the thesis, geography, company-size band, or market conditions change.

Conclusion

A PE target universe can be declared **ready for a decision**, but an unregistered fragmented market cannot be declared literally complete on the strength of a long list. The defensible sequence is to define the acquisition unit and boundary, union different candidate sources, resolve identities, review eligibility, retain source memberships, quantify incremental yield, and test overlap assumptions. A firm-count benchmark or register can anchor the answer when its scope matches the mandate. Otherwise the observed union is the strongest direct count.

Use the Chapman formula to expose how overlap changes a hypothetical total, not to turn biased company lists into a census. Compare source pairs and geography or size strata; show a conditional interval and a broader assumption range only where those calculations are meaningful. If overlap is sparse, lists share a discovery path, or an important segment is invisible to every source, say the total is unknown.

The final stopping rule is economic and strategic: another source is warranted when it is likely to add eligible companies or change the shortlist enough to justify the review cost and omission risk. Record the decision with the search date, definitions, source scopes, unresolved cases, and refresh trigger. That statement gives an origination team a reproducible basis for moving forward while preserving the uncertainty that still matters.

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/programs-surveys/susb/about.html

2. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/linkagemethodsforcensus2021inenglandandwales
3. https://www150.statcan.gc.ca/n1/pub/12-001-x/2024002/article/00007-eng.htm
4. https://applications.emro.who.int/dsaf/EMRPUB_2011_EN_1361.pdf
5. https://link.springer.com/article/10.1007/s10260-026-00840-5
6. https://arxiv.org/abs/1606.02235
7. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/coverageestimationforcensus2021inenglandandwales
8. https://arxiv.org/abs/2101.09304
9. https://ec.europa.eu/eurostat/cache/website/business-registers/metadata/european-union-business-registers-metadata-2024.htm
10. https://www.census.gov/naics
11. https://www.census.gov/programs-surveys/cbp/about/glossary.html
12. https://www.bls.gov/cew/overview.htm
13. https://unstats.un.org/unsd/business-stat/SBR/Workshops/Amman2024/Session_8_UNSD_ISIC_5.pdf
14. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/developingstandardtoolsfordatalinkagefebruary2021
15. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage
16. https://untp.unece.org/docs/specification/DigitalCorporateRecord
17. https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/methodologies/interdepartmentalbusinessregisterdatamatchinguserguide
18. https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/datalinkageandmatchingpolicy
19. https://www.sba.gov/counseling/launch-your-business
20. https://www.gov.uk/find-licences
21. https://pmc.ncbi.nlm.nih.gov/articles/PMC6055983
22. https://www.jonathantimmis.com/research
23. https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2000.00483.x
24. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04
25. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5807a3.htm
26. https://online.stat.psu.edu/stat506/Lesson12
27. https://search.r-project.org/CRAN/refmans/recapr/html/vChapman.html
28. https://hrdag.org/2013/03/20/mse-stratification-estimation
29. https://unstats.un.org/unsd/demographic/meetings/egm/symposium2001/docs/symposium_10.htm
30. https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework-guidance
31. https://unstats.un.org/unsd/business-stat/SBR/Global-Assessment/2025AssessmentResults

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.