



SCRAPPOINT · WHITE PAPER

Your B2B Database Is Not an Asset If the Data Is Dead

A practical framework for B2B data freshness and quality — the five dimensions, a database-health score, and a continuous refresh workflow.

For data managers & heads of data

Published October 2026 · Scrapoint Research
scrapoint.com

01

Executive Summary

5

dimensions of data quality

100

records to sample for a health score

3.9 yrs

median US job tenure, a decay driver

The problem. Agencies carry their prospect database on the balance sheet of their own minds as an asset — "we have 240,000 contacts." But a database is only an asset to the degree the data is true. B2B contact data decays at roughly 2.1% per month — about 22–30% a year at baseline, and faster for the volatile fields teams actually target on [1]. A database nobody refreshes is a depreciating asset that reports its original purchase price forever.

Why it matters. Dead data does not announce itself. It shows up as bounced sends, split pipeline, duplicate outreach, wasted campaign spend, and client revisions. Gartner estimates poor data quality costs the average organization \$12.9 million a year [2]; 70% of revenue leaders say they lack confidence in the accuracy of their customer data [3]. The cost is real even though the line item is invisible.

The reframe. Quality is not one thing. This paper proposes a five-dimension model — **Freshness, Completeness, Consistency, Uniqueness, Relevance** — so an agency can say how a dataset is healthy or sick, not just guess. It is Scrapoint's proposed lens, not an official standard, and its value is that it makes an invisible problem measurable.

Key findings.

1. Volume is not value: a 240,000-record database can hold far fewer usable records.
2. Decay hits hardest where it hurts most — job title, the fastest-decaying field,

is also the one agencies target on [1].

3. Duplicates are endemic: up to 80% of records entering a CRM via integrations are duplicates, and well-run operations hold the rate under 5% [4][5].

4. One-time cleanups cannot fix a continuously-decaying asset. Freshness is a process, not an event.

Recommended approach. Score the database on all five dimensions, run a quarterly sampling audit, and put collection, normalization, dedupe, and refresh in place so quality is maintained by a system rather than rediscovered by people.

02

The B2B Data Quality Problem

The core confusion is between having data and trusting it. These are different properties, and agencies routinely optimize the first while assuming the second.

Three things make B2B prospect data uniquely perishable:

- **It describes people, and people move.** Median US tenure is about 3.9 years —

the shortest since 2002 — and roughly a quarter of workers change jobs in a given year [6]. Every job change can invalidate title, company, seniority, and ICP fit at once.

- **It is entered and merged by many hands and systems.** That is how duplicates and

inconsistencies accumulate: up to 80% of records arriving through integrations are duplicates [4].

- **It is used long after it was collected.** A list built for a Q1 campaign gets

reused in Q3, by which point a predictable slice is simply wrong.

The result is a gap between the reported size of the database and its effective size — the count of records that are current, complete, consistent, unique, and still relevant. Closing the gap requires naming its parts, which is what the next six pages do.

A note on honesty before the framework: the five dimensions below are **Scrapoint's proposed working model**, offered as a practical lens for agency data teams. They are not an industry certification or an official standard, and nothing here should be presented to a client as one.

03

Dimension 1: Freshness

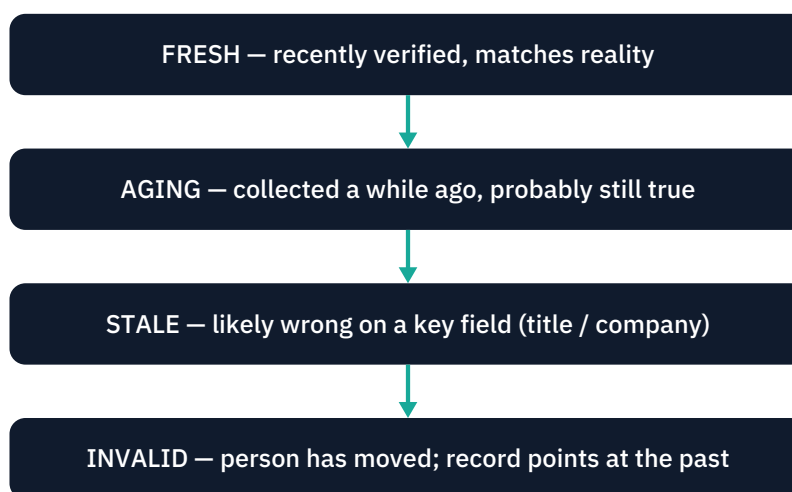
The question it answers: Is this record still true today?

Freshness is the dimension decay attacks directly. A record that was perfect on the day it was collected is not perfect six months later if the person was promoted, changed companies, or relocated. Because decay runs at roughly 2.1% per month [1], a dataset's freshness erodes continuously whether or not anyone is looking.

Why freshness is the dimension that matters most for agencies: your highest-value contacts are often the ones who move most — senior people get promoted and headhunted — so decay concentrates in exactly the records you most want to be right.

Creation date != freshness. The date a record entered your database tells you nothing about whether it is still accurate. A three-year-old record of someone who has been in the same role is fresh; a three-month-old record of someone who just changed jobs is stale. Freshness is about verification recency against reality, not age on file.

The decay progression, named:



How to improve it: re-collect current public data on a priority cadence, keyed on a stable identifier so updates land in place. (Dimensions are not improved by staring at the file; freshness specifically requires refresh — see Pages 10–11.)

04

Dimension 2: Completeness

The question it answers: Are the fields you actually use populated?

A record with a name and nothing else is not half an asset — for a campaign that segments on title and company, it is zero. Completeness is measured against the fields your work depends on, not against every field that could theoretically exist.

The trap: chasing completeness on fields you never use (a vanity of "full" records) while tolerating gaps in the fields that drive targeting. An agency that segments on seniority and function needs those parsed and present; a blank location on a record you never filter by location is not a quality problem.

A useful benchmark: well-run operations target 80%+ completion on the fields that matter [5]. The right denominator is your ICP's required fields, not the schema's total fields.

Field (example)	Used for	Counts toward completeness?
Current title	Seniority / function segment	Yes — load-bearing
Current company	Account targeting	Yes — load-bearing
Location	Geo segment (if used)	Only if you segment on it
Profile URL (key)	Dedupe + refresh	Yes — non-negotiable

Field (example)	Used for	Counts toward completeness?
Follower count	Rarely used in targeting	Usually no

How to improve it: collect the required fields at source, flag records missing any load-bearing field into an exception queue at QA, and never ship a "complete" count that is really counting empty cells.

05

Dimension 3: Consistency

The question it answers: Are the same things represented the same way?

This is the quietest killer of the five, because inconsistent data looks complete and fresh — it is just unusable in aggregate. The canonical example is the one every agency has and few measure:

As entered / collected	What a human sees	What the database sees
VP Marketing	Same role	Value A
Vice President of Marketing	Same role	Value B
VP, Marketing	Same role	Value C
V.P. — Marketing	Same role	Value D

To a person these are obviously one title. To any filter, segment, or dedupe routine they are four different strings. The consequences compound:

- A segment for "Vice President, Marketing" **silently misses three of the four**. The client's campaign is wrong and nobody sees why.
- Deduplication fails, because the matcher cannot tell that four spellings describe one role in one person's record.
- Reporting lies: the same role fragments across four buckets, so no count is right.

The fix is normalization — applied on the way in. Define a canonical form ("Vice President, Marketing"), parse seniority and function, and conform every record to it. Consistency enforced at ingestion is a property of the dataset; consistency attempted later is a cleanup project that never finishes, which is a large part of why data teams report spending the majority of their time on preparation rather than use [7].

How to improve it: one schema, canonical value forms, and a normalization stage that applies them to every record identically.

06

Dimension 4: Uniqueness

The question it answers: Is each real-world entity represented exactly once?

Duplicates are the most under-estimated quality problem in B2B data. The average company carries 10–30% duplicates, and up to 80% of records entering a CRM through integrations are duplicates; well-run operations hold the rate under 5%, which generally requires a stable key and tooling that owns dedupe continuously [4][5].

Why duplicates are expensive, not just untidy:

- They **inflate pipeline** — the database looks bigger and healthier than it is.
- They **split activity history** — one person's interactions scatter across several

rows, so no row tells the truth.

- They cause **double outreach** — two reps (or two campaigns) contact the same

person, which reads as spam and damages the relationship and the sender reputation.

The mechanism that prevents them: a stable primary key. For LinkedIn-sourced data the natural key is the **profile URL**. If every record carries it, "the same person" is one row no matter how many searches or campaigns they appeared in, and a new arrival for an existing key updates that row instead of forking a second. Without the key, deduplication is fuzzy string-matching on inconsistent names — which, per Dimension 3, is exactly the matching that fails.

How to improve it: capture the profile URL at collection, dedupe on it at write time (not in a quarterly purge), and treat uniqueness as a continuous pipeline function.

07

Dimension 5: Relevance

The question it answers: Does this record still fit the ICP?

A record can be fresh, complete, consistent, and unique — and still be worthless, because the person no longer matches who the campaign is for. Relevance is the dimension that connects data quality back to the client's actual target.

Relevance decays two ways:

- **The person changes.** A "VP Marketing at a 200-person SaaS company" who becomes a

"CMO at a 2,000-person enterprise" may have moved out of (or into) the ICP. The record is now accurate and irrelevant to that campaign.

- **The ICP changes.** The client re-scopes the target — different industry, headcount

band, seniority — and a perfectly current record is suddenly out of scope.

Because ICP fit rides on title, company, and seniority — the fast-decaying fields — relevance moves with the most volatile parts of the record [1]. This is why a dataset needs to be re-qualified against the ICP, not just re-verified for accuracy.

How to improve it: store the ICP criteria explicitly, re-score records against the current ICP during refresh, and flag records that have drifted out of scope rather than silently shipping them. Relevance is where quality meets the client's definition of a good record — and the only dimension whose standard is set outside your data team.

08 Building a Data Quality Score

With the five dimensions named, an agency can turn "I think our data is getting old" into a number it can track. The method below is **a simple internal sampling model — Scrapoint's proposed approach, not an industry-standard metric.**

The sampling audit (per dimension):

1. Pull a **random sample of 100 records** from the database.
2. For each record, check reality against the field:
 - Freshness — current title/company/location match the live profile?
 - Completeness — are the load-bearing fields populated?
 - Consistency — are values in canonical form?
 - Uniqueness — is this profile URL unique in the set?
 - Relevance — does it still match the current ICP?
3. Score each dimension = **(records that pass) / (records sampled) x 100.**
4. Average, or weight by what matters to you, for an overall health read.

Dimension	Sample question	Example score
Freshness	Does the record match the live profile today?	71 / 100
Completeness	Are the load-bearing fields populated?	88 / 100
Consistency	Are titles/locations in canonical form?	64 / 100
Uniqueness	Is the profile URL unique?	95 / 100
Relevance	Does it still fit the current ICP?	79 / 100

The individual scores are the point: an 88 on completeness and a 64 on consistency tell you to fix normalization, not to collect more. Run the same method **quarterly** and you get a trend line — the first time most agencies can see decay instead of guessing at it.

Honesty rule: label this a sampling estimate every time you report it. It is a management instrument, not a guarantee, and a 100-record sample has real margin of error. Its job is to make an invisible problem visible and trackable, not to certify a database.

09 The QA Workflow

A score tells you where the database is sick. The QA workflow is how you stop it getting sick on the way in. The division of labour is the same one that runs through a healthy pipeline: **automation handles the mechanical checks; humans handle the judgment calls.**

Automated checks (run on every record, every time):

- **Completeness** — flag any record missing a load-bearing field.
- **Consistency** — confirm normalization produced canonical values.
- **Uniqueness** — dedupe on the profile URL at write time.
- **Validity** — is the key well-formed and resolvable?

Human review (the exception queue only):

- **Relevance** calls where ICP fit is genuinely ambiguous.
- Edge cases the rules cannot adjudicate.
- Anything the client has an opinion about.

Why the split matters. Manual-only QA — one person re-reading another's work — reduces error frequency without changing error type, because both people share the same fatigue and blind spots. Automated checks change which errors are caught at all (a machine never tires of comparing profile URLs), and they free human attention for the small share of records that truly need a decision. This is also the only way QA keeps up as volume rises; a human-on-human QA layer is a linear cost that caps the operation.

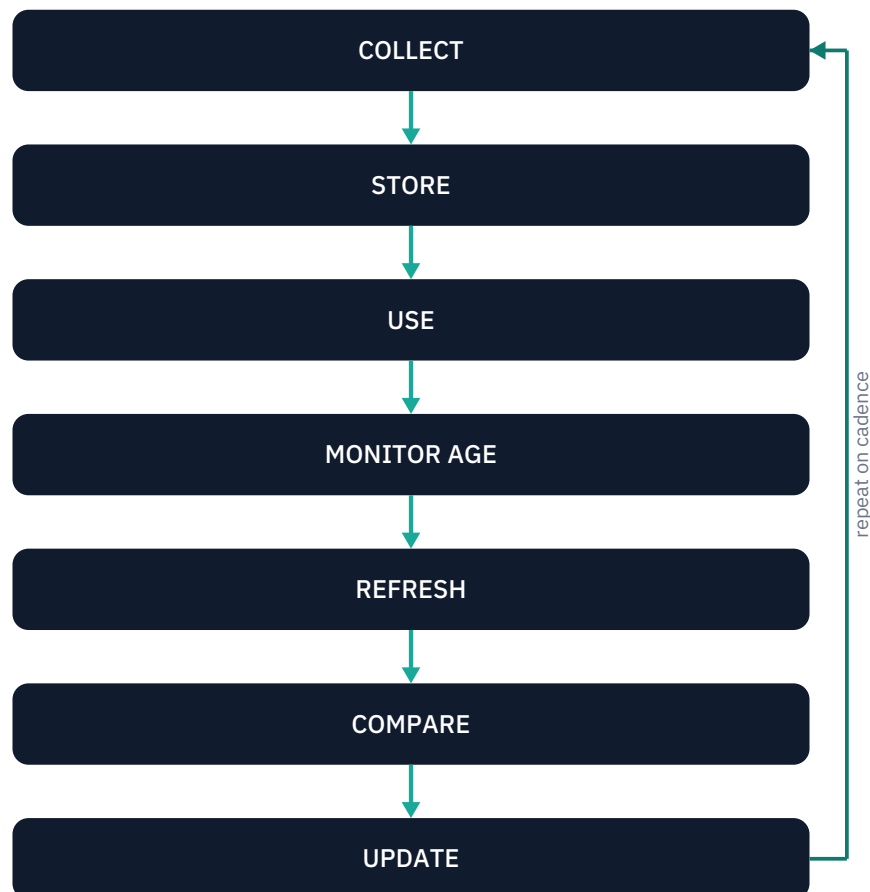
10 Continuous Data Quality

Here is the central idea of the whole paper, stated plainly:

Data cleaning is an event. Data freshness is a process.

A one-time cleanup — however thorough — fixes the database as it was on the day of the cleanup. Decay resumes the next morning at ~2.1% per month [1]. Within a year, a quarter or more of what you cleaned is wrong again. Cleaning treats quality as a destination; it is a maintained state.

The continuous model closes the loop:



- **Monitor age** — records carry a last-verified timestamp; the system knows what is getting old.

- **Refresh by priority** — value x staleness, not everything equally:

- High-value accounts — most often.
- Active-campaign contacts — before each campaign.
- Senior roles / fast industries — shorter cycle (they decay faster).
- Older, low-priority records — longer cycle.

- **Compare and update in place** — re-collect for the key, detect changed fields, update the existing row, flag exceptions.

The honest limit, again. In-place refresh works because records are keyed on a stable identifier (the profile URL). It can keep a properly-keyed store current; it cannot magically revive an arbitrary legacy database whose records have no profile URL — those must first be matched to one. Continuous quality is a property you build in, not a service you buy for a dead file.

11 The Cost of Letting Quality Slide

The five dimensions are easier to act on when tied to money. A rough model of how much of a database is actually usable, and what the rest costs:

Usable records = Total records x Freshness% x Completeness% x Uniqueness%
 Wasted spend = (Total records - Usable records) x Cost per record touched

Variable	What it is	How to get it
Freshness% / Completeness% / Uniqueness%	The dimension scores from a sample audit	The 100-record health check
Cost per record touched	Blended cost of acting on a record (rep time, sends)	Your own operating numbers

12 Worked Example: Scoring a 100-Record Sample

Illustrative — sample scores for demonstration, not measured results.

Dimension	What the sample showed	Score
Freshness	78 of 100 verified current	78
Completeness	85 of 100 have the used fields	85
Consistency	Titles / companies standardized	pass
Uniqueness	94 of 100 unique (6 duplicates)	94
Relevance	80 of 100 still fit the ICP	80

Applying the model: $100,000 \times 0.78 \times 0.85 \times 0.94 \sim 62,300$ usable records. On paper the database holds 100,000; in practice closer to 62,000 carry weight — and the gap is where wasted spend hides.

13

A 30 / 60 / 90-Day Continuous-Quality Plan

Window	Focus	Key actions	Outcome
Days 0–30	Baseline	Score a 100-record sample across the five dimensions	A quality baseline per segment
Days 31–60	Fix the source	Normalize on ingest; dedupe on the profile URL; define the used field set	Consistency and uniqueness enforced
Days 61–90	Maintain	Refresh high-value, high-volatility segments on a cadence; re-score monthly	Quality maintained as a process

14

Where Scrapoint Fits

Scrapoint maintains the first four dimensions as infrastructure and gives the fifth a place to live — stated strictly in terms of what it does:

Quality dimension	How Scrapoint supports it (factually)
Freshness	Re-collects current public data keyed on the profile URL; updates rows in place
Completeness	Collects the public profile/company/search fields into a consistent record
Consistency	Normalizes every record into one schema on the way in
Uniqueness	Deduplicates on the LinkedIn profile URL — one person, one row
Relevance	Stores records in a searchable Data Vault so segments can be re-qualified against the ICP

The boundary, stated plainly (so the framework stays honest):

- Scrapoint collects **public LinkedIn and Sales Navigator** data — profiles, company

pages, and lead/account searches — under the customer's own session and IP.

- It does **not** find or append personal **emails or phone numbers** (a different tool

category), and it does **not** invent fields a profile never published. Collection keeps the dataset in sync with what LinkedIn currently shows — which is exactly what freshness requires.

- Exports are **CSV / JSON / Excel** (and via API); pricing is credit-based, one credit

per delivered record, with a free tier of 1,000 credits.

Scrapoint is the collection + normalization + storage + refresh layer. Relevance scoring and the ICP definition remain the agency's judgment — the Data Vault's job is to make that judgment easy to apply across the whole dataset.

15 The Practical Checklist

A one-page operating checklist for a data team that wants to stop carrying a dead database as if it were an asset.

Audit (quarterly):

- ☐ Pull a random 100-record sample.
- ☐ Score Freshness, Completeness, Consistency, Uniqueness, Relevance (x / 100).
- ☐ Record the five scores with the date; build the trend line.
- ☐ Report them as sampling estimates, never as certified metrics.

Ingestion (every record):

- ☐ Capture the **profile URL** as the primary key — non-negotiable.
- ☐ Normalize titles, locations, and company references to the canonical schema.
- ☐ Dedupe on the profile URL at write time.
- ☐ Flag records missing load-bearing fields into an exception queue.

Refresh (continuous):

- ☐ Timestamp every record's last verification.
- ☐ Prioritize refresh by value x staleness.
- ☐ Re-collect for the key; compare; update in place; flag changes.
- ☐ Re-qualify refreshed records against the current ICP.

Governance:

- ☐ One source of truth (searchable store), not a drawer of spreadsheets.

- [] Automated checks for the mechanical dimensions; humans only on judgment.
- [] Keep downstream bounce under 2% as an external health signal [8].

16 Next Step

Stop reporting the size of your database and start reporting its health. A count is a vanity metric; the five-dimension score is a management instrument — and once you can see decay, you can decide to stop paying for it.

See how Scrapoint can own the collection, normalization, dedupe, and refresh that keep Freshness, Completeness, Consistency, and Uniqueness healthy by default — so your data team spends its time on Relevance and client judgment, not on resurrecting dead records.

Book a demo · scrapoint.com/product · Start free with 1,000 credits

Sources

1. Landbase & Aircscale — B2B contact-data decay rates (~2.1%/month; job title fastest):
<https://landbase.com/blog/data-decay-rate-statistics> ·
<https://airscale.io/blog/b2b-contact-data-decays-faster-than-you-think-here-is-the-evidence>
2. Gartner — poor data quality costs organizations an average of \$12.9M/year (via Data Axle):
<https://www.data-axle.com/resources/blog/stop-letting-dirty-data-drain-your-business/>
3. Cleanlist — CRM data quality benchmarks (70% of revenue leaders lack confidence in data accuracy): <https://cleanlist.ai/blog/2026-02-24-crm-data-quality-benchmarks>
4. Plauti — average rate of duplicates in the CRM (up to ~80% of integration data is duplicate):
<https://plauti.com/blog/2021-average-rate-duplicates-crm>
5. Cleanlist — CRM data quality benchmarks (duplicates <5%, 80%+ field completion):
<https://cleanlist.ai/blog/2026-02-24-crm-data-quality-benchmarks>
6. Second Talent — employee retention / tenure statistics 2026:
<https://www.secondtalent.com/resources/employee-retention-statistics/>
7. HPCwire / Anaconda — data preparation still dominates data professionals' time: <https://www.hpcwire.com/bigdatawire/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/>
8. Smartlead — cold email bounce rate & sender reputation:
<https://www.smartlead.ai/blog/cold-email-bounce-rate>