Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Survivorship Rules for Golden Records: Worked Examples

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · survivorship rules · golden record survivorship · mdm survivorship rules examples · golden record · master data best value

Original article: https://reliablecontext.com/articles/survivorship-rules-golden-record-worked-examples



In this article

1. What survivorship is for
2. The rule families
3. Worked example: one supplier, four sources
4. Where the rules disagree
5. A combined rule order with tie-breakers
6. Keeping the losing values
7. Rule governance and change control

What survivorship is for

A **golden record** is the single consolidated record a master data system publishes for one real-world entity, such as a supplier. It is built from **source records**: the rows each contributing system holds about that entity. When the source records disagree, a **survivorship rule** decides which value appears in the golden record. The glossary defines survivorship rules; this article applies the common rules to one example and shows where they disagree.

Survivorship is the last of three steps. Bleiholder and Naumann describe data integration as schema mapping, then duplicate detection, then **data fusion**, the step in which duplicate representations are "combined and fused together into a single representation while inconsistencies in the data need to be resolved" (Bleiholder and Naumann 2006, section 1). Survivorship is the master data name for that step. It assumes the records were correctly matched; no survivorship rule repairs a wrong merge.

The same authors separate two kinds of conflict (Bleiholder and Naumann 2006, section 1):

- A **contradiction**: two or more different non-null values for the same attribute of the same object.
- An **uncertainty**: one non-null value against one or more nulls. A null may mean the value is unknown, not applicable, or withheld.

Uncertainties are easy: take the value that exists. Contradictions are where rules diverge.

The rule families

Research on data fusion names the strategies (Dong and Naumann 2009, Table 2); products offer the same ones under other names. The vendor column describes product documentation, not independent evaluation.

Rule	Research name (Bleiholder and Naumann 2006)	What it picks	Vendor documentation examples
Recency	Keep up to date	The value with the latest timestamp	Microsoft "Most recent" (Microsoft Learn); Reltio LUD, its default (Reltio docs)
Source priority	Trust your friends	The value from the highest-ranked source	Microsoft "Importance"; Semarchy "Preferred Publisher" (Semarchy docs)
Completeness	Most complete (a function)	The value from the source or record with the fewest nulls	Microsoft "Most filled" (Microsoft Learn)
Frequency	Cry with the wolves	The most common value	Semarchy "Most Frequent Value"; Reltio "Frequency"
Averaging	Meet in the middle	A computed value, possibly new	Not used in the example below
Time-decayed trust	Not named there	The value whose source trust score is highest after decay	Informatica "Decay" (Informatica docs)

Deciding versus mediating. A deciding rule picks one of the values the sources supplied. A mediating rule may produce a value no source supplied, such as an average. Deciding rules preserve where-lineage, the ability to say which source a value came from; mediating rules do not (Bleiholder and Naumann 2006, section

3.2).

Avoiding versus resolving. Source priority is classed as *conflict avoidance*: whom to trust is decided once and applied whether or not a conflict exists, so "even if there is a conflict in the data, a system using these strategies does not necessarily know about it" (Bleiholder and Naumann 2006, section 2.1). Recency and frequency look at the competing values before deciding, and so are classed as resolution.

Worked example: one supplier, four sources

All names, addresses, identifiers and dates in this example are fictional. Halden Therapeutics AG keeps a supplier master. One supplier, Vossberg Labortechnik GmbH, appears in four sources:

- **REG**: an extract from the commercial register, pulled on 2026-03-02. It holds legal name and registered address only.
- **ERP**: the finance system's vendor master. Name, address and VAT ID last edited 2023-11-20; bank account changed 2026-09-15 after accounts payable confirmed it by phone on a number already on file.
- **PRC**: the procurement portal, where the supplier maintains its own profile. Most fields entered at onboarding on 2025-02-14, the address on 2025-07-08; a new bank account submitted on 2026-10-01 is unverified.
- **DP**: a purchased business-data feed. Every field carries the provider's refresh date, 2026-06-30.

Because the example is fictional, the true state is known: the supplier was renamed from Vossberg Laborbedarf GmbH in January 2025, moved from Speicherweg 3 to Kaiweg 12 in mid-2025, and its VAT ID is DE100000001.

The source values

Attribute	REG	ERP	PRC	DP
Legal name	Vossberg Labortechnik GmbH (2026-03-02)	Vossberg Laborbedarf GmbH (2023-11-20)	Vossberg Labortechnik (2025-02-14)	Vossberg Laborbedarf GmbH (2026-06-30)
Registered address	Kaiweg 12, 20457 Hamburg (2026-03-02)	Speicherweg 3, 22113 Hamburg (2023-11-20)	Kaiweg 12, 20457 Hamburg (2025-07-08)	Kaiweg 12, 20457 Hamburg (2026-06-30)
VAT ID	null	DE100000001 (2023-11-20)	DE100000001 (2025-02-14)	DE100000010 (2026-06-30)
Bank account (IBAN)	null	ending 4410, verified (2026-09-15)	ending 7781, unverified (2026-10-01)	null
Employees	null	null	85, self-declared (2025-02-14)	140, modelled estimate (2026-06-30)
Populated fields	2 of 5	4 of 5	5 of 5	4 of 5

Rule 1: most recent value

Recency takes the value with the latest timestamp per attribute. Results: legal name "Vossberg Laborbedarf GmbH" from DP (wrong, the pre-2025 name), address Kaiweg 12 from DP (right), VAT ID DE100000010 from DP (wrong, two digits transposed), IBAN ending 7781 from PRC (unverified), employees 140 from DP.

The failure is in the timestamp, not the rule. DP's date is when the provider refreshed its file, not when anyone observed the value. Recency is only as good as the date it sorts on, so keep the time a fact held, the time it was observed and the time it was loaded apart (see valid time versus transaction time). Dong and Naumann state the trade-off: "taking the most common value may end up with an out-of-date value, whereas taking the most recent value may end up with a wrong value" (Dong and Naumann 2009, section 2.3).

The IBAN result is the costly one: an unchecked portal submission displaces a bank account finance had confirmed.

Rule 2: source priority

Rank the sources once: REG, then ERP, then DP, then PRC (register first, self-reported data last). Two configurations behave differently:

- **First non-null.** Microsoft documents that its Importance option "uses the first non-null value" (Microsoft Learn). Semarchy offers a "Skip Nulls" option (Semarchy docs). Results: name and address from REG (right), VAT ID from ERP (right), IBAN ending 4410 from ERP (verified), employees 140 from DP.
- **Strict priority, nulls not skipped.** VAT ID and IBAN come out empty, because REG holds neither. An uncertainty has been turned into missing data.

Source priority wins this example because of when the extract was pulled, not because the rule is better. If the supplier moves again in November 2026, REG keeps supplying Kaiweg 12 until a new extract is pulled, and nothing flags that PRC now disagrees: the conflict-avoidance property quoted above.

Rule 3: completeness

Microsoft's deduplication offers "Most filled", which "identifies the record with most populated columns as the winner record", and breaks ties by the larger primary key (Microsoft Learn). PRC has 5 of 5 fields filled, so every value comes from PRC: name "Vossberg Labortechnik" (legal form missing), address Kaiweg 12 (right), VAT ID DE100000001 (right), IBAN ending 7781 (unverified), employees 85.

Completeness measures how much a source says, not how accurate it is. A self-service portal is often the most complete source because the supplier fills in every box, including the ones nobody has checked.

Rule 4: most frequent value

Exact-match vote: legal name "Vossberg Laborbedarf GmbH" wins 2 to 1 to 1 (wrong); address Kaiweg 12 wins 3 to 1 (right); VAT ID DE100000001 wins 2 to 1 (right); IBAN and employees are 1 to 1 ties.

Three problems show. First, normalization changes the count: strip the legal form and "Vossberg Labortechnik" joins the REG value, giving a 2 to 2 tie. Second, a tie needs a declared tie-breaker; Bleiholder and Naumann note "appropriate tie breakers are necessary" for majority strategies (2006, section 2.1). Third, votes assume

independent sources. If DP's name came from the same stale public listing that seeded ERP, the two votes are one observation counted twice. Dong, Berti-Equille and Srivastava address this case, sources that copy from others (arXiv:1503.00310).

Rule 5: reliability-weighted vote (truth discovery)

Truth discovery methods replace equal votes with source weights estimated from the data: a source that often agrees with well-supported values earns a high weight, and values backed by high-weight sources are chosen (Li et al., A Survey on Truth Discovery, section 2.2). Unlike a vote, this can select a minority value. Knowledge-Based Trust scores web sources by the correctness of the facts they state (Dong et al. 2015).

If DP had stale names and transposed identifiers on many other suppliers, its weight would fall and the REG name would win without a fixed ranking. One entity is too little data to estimate weights; these methods need many entities per source.

Where the rules disagree

Attribute	Recency	Priority (skip nulls)	Most filled	Most frequent	True state
Legal name	Laborbedarf GmbH (DP)	Labortechnik GmbH (REG)	Labortechnik (PRC)	Laborbedarf GmbH	Labortechnik GmbH
Address	Kaiweg 12	Kaiweg 12	Kaiweg 12	Kaiweg 12	Kaiweg 12
VAT ID	DE100000010	DE100000001	DE100000001	DE100000001	DE100000001
IBAN	7781, unverified	4410, verified	7781, unverified	tie	4410 until 7781 is verified
Employees	140	140	85	tie	not knowable from these sources

All four rules agree only on the address. Each fails somewhere different: recency on stale values with fresh timestamps, priority on a stale trusted source, completeness on unverified self-reported data, frequency on copied values and ties.

A combined rule order with tie-breakers

Practical survivorship is set per attribute, as ordered filters and rankings. Microsoft documents per-column merge preferences, such as "the most recent email AND the most complete address" (Microsoft Learn); Semarchy applies a rule to "a single attribute or a set of attributes" (Semarchy docs). A defensible order for the example:

1. **Eligibility filter.** Drop values that fail validation (format, check digit, verification flag). Here DP's VAT ID fails a registry lookup; PRC's IBAN waits for verification.
2. **Normalize**, then compare. "KAIWEG 12" and "Kaiweg 12" are one value; so are two names that differ only in how the legal form is written.

3. **Attribute-specific ranking.** Legal name and registered address: REG first, then the remaining sources by frequency. VAT ID: REG, ERP, PRC, DP. IBAN: verified values only, most recent first. Employees: self-declared value if under 24 months old, else provider estimate.
4. **Tie-breakers:** source rank, then observation date, then lowest source record ID, so every run returns the same result.
5. **Group dependent fields.** Street, postcode and city survive together. Microsoft's documentation shows why: without a group, per-field ranking combined fields from two different records into an address that exists in neither (Microsoft Learn).
6. **Escalate.** When the filter removes a value newer than the survivor, open a review task.

Attribute	Golden value	Deciding step	Losing values kept
Legal name	Vossberg Labortechnik GmbH	REG rank	Laborbedarf GmbH (ERP, DP); Labortechnik (PRC)
Registered address	Kaiweg 12, 20457 Hamburg	REG rank	Speicherweg 3 (ERP, 2023-11-20)
VAT ID	DE100000001	Filter, then ERP rank	DE100000010 (DP, failed lookup)
IBAN	ending 4410	Verified-only filter	ending 7781 (PRC, pending verification, review open)
Employees	85	Self-declared under 24 months	140 (DP estimate)

The combined order gets every knowable attribute right, but it was written by someone who knew the answers. That is the main limitation of rule design: rules are fitted to failures already seen.

Keeping the losing values

Every rule above discards information, and the discarded value is often the one an auditor needs. The PRC bank account may be legitimate but unverified; the DP employee count may be the better figure next year. If only survivors are stored, nobody can later say what else the sources said and why it was rejected.

Store, for each attribute: every candidate value, its source, the date it was observed as distinct from the date it was loaded, the rule version that decided, and the deciding step. This is cheap with deciding rules, since each survivor is one of the inputs and keeps where-lineage (Bleiholder and Naumann 2006, section 3.2). It also matters for maintenance: Dong and Naumann note that non-associative functions such as voting or averaging can give incorrect results when new conflicting values arrive, and name retaining data lineage among the techniques needed (Dong and Naumann 2009, section 2.5).

Products differ in what they keep. Microsoft states that non-winner rows in deduplication are still used in matching (Microsoft Learn); Informatica's SaaS documentation states that "data from different source systems is stored in source records" (Informatica docs). Check what your product retains and whether a reviewer can

see it beside the golden value. The case for publishing the preferred value with its evidence is made in golden record versus evidence-carrying record; the procedure for writing up an unresolved conflict is in reporting source conflicts.

Rule governance and change control

Survivorship rules change published data, so treat them like code:

- **Version every rule** and record the version with each decided value, so an earlier golden record can be reproduced from retained source values.
- **Know your product's recalculation behaviour.** Informatica's Business 360 documentation states that after a survivorship configuration change, "trust scores of the business entity records aren't recalculated", and are recalculated only when users edit the records (Informatica docs). Old and new rules then coexist until each record is touched.
- **Make decay explicit.** Informatica's decay rule sets a maximum trust score between 0 and 1, a minimum, a decay period and a pattern (linear, rapid then slow, or slow then rapid) (Informatica docs). These parameters claim how fast each source goes stale; review them against observed errors.
- **Replay before release.** Run a proposed rule on a frozen sample and list every golden value that would change.
- **Watch the escalation queue.** A growing number of filtered-out newer values usually means a source changed, not that the rule is working.

A golden record is easiest to trust when a reader can see, for any value, which rule chose it and which values it set aside.

Source links

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

1. https://edoc.hu-berlin.de/18452/3112
2. https://lunadong.com/publication/fusion_vldbTutorial.pdf
3. https://learn.microsoft.com/en-us/dynamics365/customer-insights/data/data-unification-merge-tables
4. https://docs.reltio.com/objectives/resolve-potential-matches/potential-matching-at-a-glance/potential-matching-navigation/design-survivorship-rules/configure-default-survivorship-strategy
5. https://www.semarchy.com/doc/semarchy-xdm/xdm/2024.4/Design/matching/create-a-survivorship-rule.html
6. https://learn.microsoft.com/en-us/dynamics365/customer-insights/data/data-unification-duplicates
7. https://onlinehelp.informatica.com/IICS/dev/b360/en/gg-b360-configure-survivorship/Rule.html
8. https://arxiv.org/abs/1503.00310
9. https://arxiv.org/abs/1505.02463
10. https://arxiv.org/abs/1502.03519
11. https://onlinehelp.informatica.com/IICS/prod/mde/en/cg-c360-bvt-records/Establish_the_best_version_of_the_truth.html
12. https://onlinehelp.informatica.com/IICS/dev/b360/en/gg-b360-configure-survivorship/Editing_survivorship_rules.html

13. https://onlinehelp.informatica.com/IICS/dev/b360/en/gg-b360-configure-survivorship/Step_3_Configure_survivorship_rules.html

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

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.