

Duplicate Checking: Clean Master Data as the Foundation of Procurement

Duplicate checking refers to the systematic identification and cleansing of duplicate records in master data inventories, in particular supplier and material master data. It is a core data management task in procurement, as duplicates can lead to duplicate purchase orders, incorrect payments and distorted analyses. In practice, the check is carried out both manually and with software tools and is repeated on a regular basis.

Key facts

- Duplicate checking identifies duplicate supplier, material and customer master records in the system.
- Duplicates typically cause incorrect purchase orders, double payments and distorted reports.
- The check is carried out using exact, fuzzy and context-supported matching techniques.
- Regular cleansing improves data quality KPIs such as record quality and the invoice error rate.
- A clean supplier base is a prerequisite for supplier segmentation and reliable consolidation levers.

Definition: duplicate checking

Duplicate checking means checking data inventories for identical or substantively identical records that have been created multiple times under different keys. It forms part of master data management and is frequently carried out ahead of migrations, S/4 conversions or system consolidations. The result is a cleansed, consistent data foundation with unique references.

Distinction from data maintenance

Duplicate checking is a sub-area of ongoing data maintenance: it focuses on detecting multiple record creations, whereas general maintenance also covers classification, field histories and authorisations. The two interact, because without clear creation processes, duplicates reappear. In practice, companies therefore establish check workflows before new master records are created.

Affected data objects

Typically, duplicates affect three object classes: supplier master records (the same service provider under several numbers), material master records (the identical part with several material numbers) and open purchase orders. Supplier duplicates are particularly critical because they obscure consolidation levers in procurement and distort the supplier segmentation concept. Bank details and addresses can also exist multiple times per supplier.

Methods and approach in duplicate checking

The approach usually follows a multi-stage pattern: data preparation, potential matching, review of candidates and consolidation. First, field contents are normalised, i.e. capitalisation, umlauts, legal forms and address formats are standardised. Matching then takes place across defined attribute

combinations, followed by a business assessment by experts.

Exact matching techniques

In the exact check, identical values are searched for in selected fields, for example identical tax numbers, VAT identification numbers, IBAN/BIC-SWIFT combinations or supplier numbers from legacy systems. This method delivers reliable hits with a low checking effort. However, it skips cases where spelling variants or address deviations exist, which is why it usually serves only as a first stage.

Fuzzy matching and similarity searches

Fuzzy techniques score similarities based on algorithms such as Levenshtein distance or phonetic search and find hits such as “Müller GmbH” and “Mueller GmbH”. In practice, similarity thresholds are set so that a manageable candidate pool is produced, typically a few per cent of the total data volume. The final decision is always made by a responsible expert reviewer.

Context-supported checking

In addition, business data can be used: the same order addresses, identical contacts, matching delivering plants or similar invoice patterns. Payment data such as identical bank account details are also strong indicators. This contextual view complements the field comparisons and uncovers cases that pure text comparisons overlook.

Key KPIs of duplicate checking

The effectiveness of duplicate checking can be tracked using metrics that are reported regularly as data quality KPIs. They show both the cleansing progress and the effectiveness of preventive measures. Typical benchmarks help in assessing your own level of maturity.

Duplicate rate and record quality

The duplicate rate indicates the proportion of identified duplicates relative to the total volume; in practice, uncleansed legacy records frequently show values between 2 and 10 per cent. Record quality measures the proportion of records that are complete and consistently maintained. Both values should be documented after each cleansing round.

Process-related metrics

Also relevant are the invoice error rate, since supplier duplicates lead to multiple allocations, as well as the number of new record creations rejected by the creation workflow. The average processing time per duplicate case is also a common steering metric. These KPIs show whether root causes are being addressed or only symptoms treated.

Risks and controls relating to duplicates

Duplicates are more than a nuisance: they cause measurable costs and increase compliance risks. Double payments to separate supplier numbers, multiple purchase orders for identical material and incomplete supplier evaluations are typical consequences. Controls therefore belong in both the system architecture and the processes.

Financial and quality-related consequences

Duplicates make total cost accounting more difficult, because spend and volume analyses become

fragmented and consolidation potential remains undiscovered. In inventory management, material duplicates lead to double stock levels and incorrect stock analyses. In accounts payable, the risk of double payments increases, making retrospective archive audits expensive.

Organisational controls

Effective controls start at record creation: mandatory fields, plausibility checks, a defined creation workflow with approval and a central role for master data maintenance. For suppliers, checks against the Approved Vendor List (AVL) and reconciliation of bank details supplement the controls. Periodic re-audits of the data basis maintain quality on a lasting basis.

Compliance aspects

With supplier duplicates, there is a risk that compliance checks only capture one of the two records, for example in sanctions list screening or documentation of child labour risk assessments in the supply chain. Requirements arising from Intrastat reporting also presuppose consistent supplier and material master records. Duplicate checking should therefore be firmly anchored in the compliance calendar.

Current developments and automation

Duplicate checking is becoming increasingly automated: modern systems identify potential duplicates in real time at the point of creation rather than only after the fact. Artificial intelligence and machine learning models improve match accuracy in fuzzy comparisons and noticeably reduce the manual checking effort.

Real-time checking at record creation

Instead of periodic batch runs, modern ERP and master data solutions already check in the creation form whether a similar record exists. The creator receives immediate hints and can avoid duplicates from the outset. In practice, this preventive logic is regarded as the most effective measure because it prevents errors at the source.

Integration into procurement processes

Cleanly resolved master data has an effect beyond duplicate checking: it improves capacity planning in procurement logistics, facilitates purchase order release and makes tenders and e-auction preparation more reliable. Buyers also benefit from unambiguous material identification in inventory management and demand consolidation.

AI-supported classification

Language models and pattern recognition help to normalise free-text fields and automatically compare material short texts. This also makes it possible to identify duplicates that differ only in wording, not in key. The decision on whether to merge records nonetheless remains with the responsible expert reviewer.

Typical metrics

Duplicate rate: < 1 %

Target value after cleansing; uncleansed legacy records typically range from 2–10%.

Record quality: > 95 %

Proportion of complete, consistent master records; a common benchmark for well-maintained data inventories.

Invoice error rate: < 2 %

Proven to fall once supplier duplicates have been eliminated.

Duplicates per new record creation: > 98 % prevention rate

Proportion of duplicate creation attempts intercepted by real-time checking at the point of creation.

Practical example

A mid-sized mechanical engineering company with around 12,000 active suppliers and 80,000 material master records carried out a comprehensive duplicate check ahead of a system migration. Over several iterations, exact matches on VAT ID numbers and bank details were combined with fuzzy searches on company names and addresses.

- Establishment of a creation workflow with real-time duplicate warnings for all new supplier and material master records
- Consolidation of around 900 supplier duplicates, followed by reconciliation against the Approved Vendor List (AVL)
- Cleansing of material duplicates with a subsequent stock analysis to merge double stock levels
- Introduction of quarterly data quality KPIs for the ongoing monitoring of the duplicate rate

Conclusion

Duplicate checking is an seemingly unglamorous yet economically relevant instrument of master data management. It safeguards the validity of spend analyses, reduces the risk of double payments and creates the data foundation for consolidation, supplier segmentation and reliable metrics. Those who organise the check not as a one-off project but as a continuous process with clear accountability prevent duplicate record creation at the source. Modern real-time checks and AI techniques considerably reduce the effort involved. Investment in clean master data pays off directly in almost all downstream procurement processes.

Frequently asked questions

What is duplicate checking?

Duplicate checking is the systematic search for duplicate records in master data inventories, for example among suppliers or materials. It combines exact field comparisons with fuzzy matching techniques. The aim is a consistent data foundation with unique references.

How do you identify duplicates in supplier master data?

Reliable indicators are identical VAT identification numbers, tax numbers, bank details (IBAN/BIC-SWIFT) and company addresses. In addition, fuzzy matching techniques are used that detect spelling variants and umlaut deviations. The final business review of the candidate list is carried out by the responsible staff.

What costs do duplicates cause in procurement?

Duplicates lead to duplicate purchase orders, double payments and fragmented spend analyses, meaning that consolidation potential is lost. Material duplicates cause unnecessary double stock levels

in inventory management. In practice, these follow-on costs significantly exceed the effort required for cleansing.

How often should duplicate checking be carried out?

A proven approach is a two-stage rhythm: a real-time check on every new record created in the system and a periodic full check, typically quarterly or half-yearly. A comprehensive cleansing is mandatory ahead of migrations or system consolidations. The frequency depends on data growth and the intensity of changes.

Which KPIs measure the success of duplicate checking?

Key metrics are the duplicate rate, the record quality and the development of the invoice error rate. In addition, the number of rejected new record creations and the processing time per case are reported. These data quality KPIs show whether preventive measures are taking effect.