

Duplicate score: key figure for master data quality in procurement

The duplicate score is a key figure that measures how free master data sets – such as material, supplier or item masters – are from duplicate entries. It is typically expressed as a percentage and is a central indicator of data quality in master data management. It is relevant for buyers because duplicates can lead to double stock-keeping, incorrect price comparisons and distorted analyses.

Key facts

- The duplicate score sets the number of unique records in relation to the total number of records.
- In practice, a good value is often above 95 per cent; below this, the cleansing effort increases significantly.
- Duplicates distort spend analyses, price comparisons and the Purchase Price Variance (PPV).
- Determination is usually carried out using fuzzy matching, i.e. approximate similarity searches in names, numbers and attributes.
- A low score typically increases the review effort in the invoice approval workflow.
- Regular cleansing is part of functioning master data management and reduces duplicate orders.

Definition: duplicate score

The duplicate score quantifies the proportion of duplicate or potentially duplicate records in a master data set. It serves as a measure of data quality and as a control variable for data cleansing projects.

Basis of calculation

In practice, the score is often calculated as the proportion of records classified as unique out of all records. Alternatively, companies count the identified duplicate candidates and set them in relation to the total stock. It is crucial that the definition is documented consistently within the company so that values remain comparable across periods.

Distinction from the error rate

The duplicate score only covers one quality dimension: redundancy. Other errors such as incomplete attributes, incorrect units of measure or outdated supplier assignments are recorded separately. For holistic master data management, the score should therefore be combined with further data quality key figures.

Methods and approach for duplicate identification

Determining the duplicate score follows a multi-stage process: profiling, matching, classification and cleansing. The score is measured both before and after cleansing in order to demonstrate progress.

Matching methods

Exact matches on numbers are rarely sufficient, as duplicates often arise through typos, deviating descriptions or different units of measure. Fuzzy matching with similarity scores is therefore common, supplemented by match-and-merge rules that define the degree of similarity at which two records are considered duplicates. Attributes such as dimensions, materials and standards serve as confirmation criteria.

Classification and approval

In practice, identified candidates are rarely merged automatically. A match usually undergoes manual review by the specialist departments, similar to a request for additional documents in supplier management. Only after approval are records merged, whereby histories such as orders and prices must be preserved.

Key KPIs relating to the duplicate score

In practice, the duplicate score never stands alone. It is combined with supplementary key figures in order to draw conclusions about costs, process quality and data reliability.

Data quality key figures

In addition to the score, completeness levels, the currency of master data and the number of records created per period are relevant. A rising rate of new records combined with a declining score often indicates that search functions or creation processes are inadequate.

Impact on procurement key figures

Duplicates distort spend bundling and thus potential economies of scale in supplier consolidation. The Purchase Price Variance (PPV) can also appear distorted if prices are posted to different records for the same material. The score therefore also serves as a stress test for key figures in the procurement business plan.

Risks and controls

An undetected poor duplicate score has an impact along the entire procurement chain. Controls should therefore be integrated into processes rather than carried out as one-off exercises.

Operational risks

Typical consequences are double stock and orders, incorrect framework agreements and multiple payments to the same supplier under slightly deviating names. In the invoice approval workflow, the manual clarification effort increases when invoices match different records for the same transaction.

Strategic risks

At a strategic level, duplicates distort the supplier portfolio: spare parts are procured individually instead of being bundled, and supplier evaluations – for example as part of pre-qualification or supplier certification – are based on incomplete data. Clean master data is also a prerequisite for audits and export control.

Control measures

Duplicate checks at the point of record creation, periodic score measurements with trend tracking and

clearly defined responsibilities in master data management have proven effective. In addition, procurement policies – such as an MRO vs. Capex procurement policy – govern who may create material masters and according to which criteria they are checked.

Current developments: automation and data maintenance

AI-supported classification and semantic search significantly improve duplicate detection rates, particularly for free-text item numbers. At the same time, the manual effort required for review and merging decreases.

AI and semantic matching

Modern methods can identify duplicates even where descriptions differ considerably, for example between manufacturer and plant designations. In practice, however, software suggestions are usually confirmed by specialist departments in order to avoid incorrect merges, which are difficult to reverse.

Integration with procurement processes

Cleanly prepared master data is a prerequisite for automated processes such as Available to Promise (ATP) in availability checks or consistent prices including the Raw Material Surcharge. In the ramp-up management of new products, too, a high duplicate score contributes directly to schedule adherence, because spare part and material data can be reliably assigned.

Typical metrics

Duplicate score: > 95 %

typical benchmark for a solidly cleansed master data set

Rate of new records created as duplicates: < 2 %

proportion of new records later identified as duplicates

Cleansing rate per quarter: > 90 %

proportion of confirmed duplicate candidates actually merged

Practical example

A medium-sized mechanical engineering company with around 120,000 material master records discovered during a spend analysis that several standard parts were being maintained under different descriptions. The initial duplicate score was in the region of 92 per cent unique records. After a clean-up using fuzzy matching and review by the specialist department, the value stabilised over several quarters at above 97 per cent.

- Introduction of a mandatory search before creating new material master records
- Definition of company-wide match-and-merge rules with approval levels
- Quarterly score measurement with reporting to procurement and IT
- Consolidation of standard parts across fewer suppliers to achieve economies of scale

Conclusion

The duplicate score is a simple yet effective key figure for the quality of procurement and material master data. It makes redundancy visible, which otherwise generates hidden costs – through duplicate

orders, distorted analyses and increased review effort. What matters is a consistent calculation method and regular measurement, not a one-off cleansing run. Combined with clear creation processes and defined responsibilities, the score contributes directly to procurement optimisation. For technical buyers, it is thus a reliable early indicator of the robustness of all data-based decisions.

Frequently asked questions

What is a duplicate score?

The duplicate score is a key figure indicating the proportion of duplicate records in a master data set, for example in the material or supplier master. It is typically expressed as a percentage. A high score means a low number of duplicates and therefore good data quality.

How is the duplicate score calculated?

The usual formula is: unique records divided by the total number of records, multiplied by 100. Identification of duplicates is typically carried out using fuzzy matching and defined match-and-merge rules. It is important to document the methodology so that values remain comparable across periods.

From what level is a duplicate score considered good?

In practice, values above 95 per cent are considered solid and above 98 per cent very good. However, the benchmark depends on the industry and the data set. More important than the absolute value is the trend: a declining score points to problems in the record creation process.

What costs do duplicates cause in procurement?

Duplicates lead to double stock-keeping, missed bundling effects and additional effort in the invoice approval workflow. They also distort spend analyses and the Purchase Price Variance (PPV). The actual cost impact varies from company to company and can only be determined on a case-by-case basis.

How can duplicates in master data be reduced permanently?

Prevention is key: mandatory search before creating new records, standardised naming conventions and defined creation authorisations. In addition, periodic duplicate runs with a clear approval process help. One-off large-scale clean-ups without process changes frequently lead in practice to a renewed rise in duplicates.