

Duplicate Detection: Systematically Identifying Duplicate Records in Master Data and Ordering Processes

Duplicate detection refers to the systematic identification of duplicate records in material master data, supplier, customer and order data. It is a core discipline of data quality management in industrial procurement, as duplicates lead to duplicate orders, price fragmentation and incorrect analyses. In practice, a considerable share of duplicates arises from decentralised material creation, inconsistent descriptions and a lack of standardisation. Effective duplicate detection lowers process costs and creates a reliable data foundation for tenders, catalog ordering and supplier consolidation.

Key facts

- Duplicate detection identifies records that are substantively identical but formally different in master and transaction data.
- Typical causes are decentralised material creation, inconsistent descriptions and a lack of classification.
- Proven methods include exact key comparisons, fuzzy matching and clustering, often combined with AI-supported text similarity analysis.
- In practice, the share of detected duplicates in the material master is often in the low single-digit percentage range, but can reach several per cent.
- Cleansed master data is a prerequisite for reliable data quality scores, spend analyses and catalogue management.
- Duplicate prevention via a creation-gating workflow is usually more effective than subsequent cleansing.

Definition: Duplicate Detection

Duplicate detection is the process by which records are found in datasets that describe the same real-world object — for instance the same material, the same supplier or the same order — even though they have different keys, designations or spellings. It encompasses the technical methods of similarity checking as well as the organisational steps for reviewing, merging and cleansing the identified entries. In procurement, it focuses primarily on material master data, supplier master data, order line items and invoice data.

Distinction from Data Cleansing

Duplicate detection is one step within the broader process of data cleansing, which also covers incomplete, outdated or contradictory data. While cleansing addresses many data errors at once, duplicate detection concentrates on the multiple recording of the same object. Together, both feed into a meaningful data quality score with which companies can make the state of their master data

measurable.

Why Duplicates Are Particularly Costly in Procurement

Duplicates in the material master do not cause visible errors but hidden costs: the same part is procured under several numbers, volume bundles remain undiscovered and supplier terms fragment. Duplicate supplier master records also complicate compliance processes such as KYC in procurement, as risk assessments are made on incomplete data bases. Duplicates likewise create effort and misallocation for Intrastat reporting and tender management processes.

Causes and Origins of Duplicates

Duplicates rarely arise intentionally, but mostly from structural weaknesses in data creation. Those who know the causes can derive targeted creation-gating measures and permanently reduce the duplicate rate.

Decentralised Creation and Missing Check Loops

When materials are created across several plants, sites or systems without central checking, multiple entries of the same part quickly arise. If a creation-gating workflow with automated pre-check is lacking, each organisational unit only checks locally whether the record already exists. The result is a growing stock of "shadow materials" that map the same real-world item.

Free Text Instead of Standardised Descriptions

"Flange DN80 PN16 stainless steel", "Flange, DIN, 80mm, VA" and "FLANGE ST37 DN80" may describe the same part. Without standardisation and classification according to a standardised scheme, the similarity remains invisible to simple searches. Missing attribute data — for instance on dimensions, material or standard — also prevents unambiguous comparison across system boundaries.

Mergers, System Changes and Supplier Master Data

System migrations and site mergers typically cause surges in duplicates, because inventories from different systems are merged without reconciliation. In supplier master data, duplicates arise from different spellings, site assignments or tax and legal variants of the same corporate group. Such duplicates impair supplier consolidation and distort the assessment of the A-supplier portfolio.

Methods and Approach of Duplicate Detection

The choice of method depends on the data volume, the structure of the data and the desired hit accuracy. In practice, a combination of simple rule-based checks and more flexible similarity-based methods has proven effective.

Rule-Based and Deterministic Methods

Deterministic methods compare exact keys: manufacturer part number plus plant number, identical EAN, or a combination of standard, dimension and material. These checks are fast, transparent and generate hardly any false positives. Their drawback: they overlook duplicates where spellings, abbreviations or units differ — which is the norm in free-text master data.

Fuzzy Matching and Clustering

Fuzzy matching methods evaluate the string similarity of designations and attributes and also find

duplicates with slightly deviating spellings. Clustering methods group records that probably represent the same object without requiring an exact reference record. Modern approaches use language models or semantic similarity to also recognise synonymous descriptions; the price is a higher manual review effort for the matches.

Approach in a Cleansing Project

A typical approach consists of five steps: data profiling and definition of match rules, automated pre-run over the inventory, manual review of candidates by domain experts, merging of the records with reference history and subsequent creation-gating lock so that duplicates do not build up anew. The match classes should be tiered by confidence so that unambiguous duplicates can be consolidated automatically and unclear cases are reviewed.

Key KPIs and Measuring Success

Without measurement, duplicate management remains invisible work. Proven metrics make inventory quality, progress and prevention measurable and support the business case towards management.

Duplicate Rate and Data Quality Score

The duplicate rate — the share of detected duplicates among all master records — is the core metric. In practice, it is often in the range of a few per cent for uncleansed material masters; after cleansing, values close to zero are targeted. The data quality score combines the duplicate rate with further criteria such as completeness, currency and degree of classification into an overall assessment of data quality.

Process Metrics and Drift Control

Also important are the processing speed of the review stacks (e.g. days to decision), the share of automatically consolidated cases and the new-record rate with duplicate incidents. A sensible control is the monthly measurement of newly arising duplicates: if this value remains constantly low, the creation-gating measures are working. The order rate per catalog serves as an indirect indicator, because a clean material basis facilitates catalog assignment and reduces free-text enquiries.

Risks, Controls and Organisation

Duplicate detection carries its own risks: incorrectly merged records can distort supplier relationships, order histories and stock levels. Controls and clear responsibilities are therefore indispensable.

Risk of Incorrect Consolidation

Two similar records need not be identical: material variants, tolerances or certificate requirements can technically change a procurement. A premature merge can lead to the wrong part being ordered or alternative procurement via duplicate candidates simply being forgotten. Therefore every merge requires an expert review and a documented approval — for materials ideally by engineering or quality assurance.

Data Governance and Role Model

Effective duplicate prevention requires clear data governance: who may create materials, who reviews, who approves consolidations? A central master data management function with defined classification rules and mandatory fields has proven its worth. In addition, ordering processes should be designed so that order placement is only possible on approved master records — this reduces the incentive to use

unchecked duplicate entries.

Prevention Through Catalogues and Procurement Policies

Catalog buying significantly reduces duplicate risks, because procurement runs via validated catalog items instead of free-text new entries. When demands consistently run through requirement confirmation and approved catalogues, the main incentive for duplicate entries disappears. In addition, demand aggregation and spend consolidation help make multiple demands visible before they lead to parallel orders.

Current Developments and Outlook

Duplicate detection is evolving from a one-off cleansing project into a continuous process. Drivers include AI-based text analysis, growing compliance requirements and the digitalisation of procurement.

AI and Semantic Similarity Search

Modern methods use vector-based similarity search and large language models to recognise even strongly deviating descriptions of the same part. This opens up the possibility of suggesting duplicates in real time already in the creation workflow — the user sees "may already exist" rather than retrospectively generated review stacks. In practice, the balancing act remains between a high hit rate and burdening specialist users with review effort.

Compliance and Supplier Data

Increasing regulatory pressure, for example in the context of the Corporate Sustainability Due Diligence (CSDDD), makes consistent supplier master data a mandatory task: risk assessments and due diligence checks presuppose that the same supplier relationship is not maintained under three different names. KYC in procurement also benefits from unambiguous supplier identities. Duplicate detection is thus increasingly becoming part of the compliance architecture, not just cost reduction.

Integration into Procurement Platforms

In future, duplicate checks will increasingly be integrated into procurement platforms, marketplaces and ERP systems rather than run as isolated point-in-time projects. Synchronisation with supplier catalogues, automatic classification suggestions and reconciliation with external manufacturer data help ensure that new entries are clean from the start. Change order clauses and order change processes also benefit, because consolidated histories make the sequence of changes traceable.

Typical metrics

Material master duplicate rate: typically 1–5% before cleansing

Proportion of detected duplicates among all master records; target value after cleansing close to 0%

Data quality score: frequently target value > 95%

combines duplicate rate, completeness and degree of classification

New duplicate rate per month: typically < 0.1%

measures the effectiveness of creation-gating prevention

Order rate per catalog: rising value

indirect indicator, as clean master data increases catalog assignment

Practical Example

A mid-sized mechanical engineering company with several plants discovered during a spend analysis that standard parts such as standard threaded fittings and bearings were being procured under up to five different material numbers. The duplicate rate in the material master was in the low single-digit percentage range after an initial automated analysis, with considerable price fragmentation potential for frequently procured standard parts. After a six-month cleansing and prevention programme, the company was able to consolidate its volume bundles and significantly reduce the number of active material numbers.

- Automated pre-run over 250,000 material master records with fuzzy matching and clustering of candidates
- Expert review of match stacks by engineering and procurement, with documented approvals for consolidation
- Creation-gating workflow with real-time similarity warning in the material management system
- Binding classification policy with mandatory fields and migration of standard parts to catalog buying

Conclusion

Duplicate detection is far more than a technical cleansing tool: it creates the data foundation on which every reliable analysis, tender and supplier management activity is built. Those who only cleanse duplicates retrospectively are fighting symptoms — what matters is creation-gating prevention through workflows, classification rules and catalog procurement. The investment typically pays for itself through better terms, lower process costs and reliable compliance data. Companies should organise duplicate detection as a continuous process, not a one-off project. This turns a data quality problem into a measurable competitive advantage in procurement.

Frequently asked questions

What is duplicate detection?

Duplicate detection is the systematic identification of data records that describe the same real-world object but have different keys or spellings. In procurement, it primarily affects material, supplier and order data. It uses both exact key comparisons and similarity-based methods.

How do you detect duplicates in the material master?

Initially through rule-based comparisons of characteristics such as manufacturer part number, standard and dimensions. For free-text descriptions, fuzzy matching, clustering and increasingly AI-based semantic similarity search are used. The candidates then need to be reviewed by domain experts, as similar records are not automatically identical.

What costs do duplicates cause in procurement?

Duplicates lead to duplicate orders, fragmented supplier terms and unused volume bundling. In addition, process costs rise due to redundant part numbers, stock levels and search effort. No universal figure can be given, but in practice the potential for standard and repeat parts is considered particularly significant.

What is a typical duplicate rate in the material master?

In uncleansed, decentrally maintained master data, the detected duplicate rate is often in the low

single-digit percentage range, in extreme cases even higher. After a clean-up, companies aim for values close to zero. What matters less is the starting value than the control of newly emerging duplicates.

What is the difference between duplicate detection and data cleansing?

Data cleansing covers all measures to improve data quality, including completion, updating and correction. Duplicate detection is the sub-area that focuses on identifying and consolidating multiple entries of the same object. Both feed into a data quality score.

How can duplicates be prevented permanently?

What matters is prevention at record creation: creation-gating workflows with automatic similarity warnings, binding classification and mandatory field rules, and central master data ownership. In addition, catalog buying reduces free-text new entries because demands are processed via validated catalog items. Regular monitoring runs sustain the state over the long term.