

5th TOLERANT Whitepaper

How to find and avoid duplicates in the database

Best practices and case studies

Chapters 5-6 from the book entitled "Wie man Dubletten im Datenbestand findet und vermeidet" by Klaus-Dieter Sedlacek. ISBN 978-3-384-20706-7 (Reprinted with kind permission of the author)



TOLERANT Software

1. Best practices and guidelines

1.1 Strategies for the ongoing maintenance and updating of person search fields

The ongoing maintenance and updating of people search fields is an essential process for organizations to ensure the integrity, reliability and usability of their data assets. Given the dynamic nature of personal information - people change names, addresses, phone numbers and other relevant data - data maintenance requires ongoing attention and the application of best practices. Here are some best practices and guidelines that can help organizations effectively manage their people search fields:

1. **Regular data checks:**

Carry out systematic checks of the databases at regular intervals to identify inaccuracies, obsolescence and possible duplicates. This can be done every six months or annually, depending on the scope of the data and the frequency of changes.

2. **Automation of data updating:**

Use automated tools and software solutions that support the detection of changes in person search fields. For example, data matching tools can help to identify new information and update existing records accordingly.

3. **Feedback loops with users:**

Allow users or customers to check their own data and suggest corrections or updates. This can be done via online portals or through regular requests to check the data.

4. **Employee training:**

Ensure that your employees are trained in the correct procedures for entering and updating data. This includes guidelines on formatting, dealing with special cases and avoiding data entry errors.

5. **Introduction of data quality standards:**

Define clear standards and guidelines for the quality of person search fields. This includes specifications for formatting, data completeness and the use of standard codes (e.g. country codes).

6. **Use of data cleansing services:**

Consider using external data cleansing and enrichment services to improve the timeliness and accuracy of people search fields. Such services can help validate addresses, update phone numbers and check email addresses for validity.

7. **Data protection and compliance:**

Ensure that all data maintenance and updating measures comply with applicable data protection laws and regulations. This includes the secure handling of personal data and ensuring transparency and control for data subjects.

Adhering to these best practices and guidelines enables organizations to manage their people search fields efficiently and responsibly. Continuously maintaining and updating this data not only helps to improve data quality, but also supports data-driven decision making, optimizes business processes and promotes user confidence in the organization's data processing practices.



1.2 Quality control and testing of preparation methods

Quality control and testing of preparation methods are essential parts of the data cleansing process to ensure that the methods used are effective in improving data quality. These steps serve to check and validate the consistency, accuracy and reliability of the processed data before it is used for further analysis, reporting or operational processes. Here are key aspects and guidelines for performing effective quality control and testing of preparation methods:

1. **Development of quality standards:**
Define clear and measurable quality standards for the processed data. These standards should focus on the accuracy, completeness, consistency and timeliness of the data.
2. **Implementation of validation rules:**
Use automated validation rules to check compliance with the specified quality standards. This can include checking formatting, compliance with value ranges and the consistency of key information.
3. **Test data sets:**
Create representative test data sets that cover a wide range of scenarios and potential data errors. These data sets should be used to test and evaluate the effectiveness of the preparation methods.
4. **Carrying out test runs:**
Carry out test runs with the preparation methods to check their functionality and effectiveness under different conditions. Analyze the results to identify areas for improvement.
5. **Benchmarking and performance comparisons:**
Compare the performance of the preparation methods against defined benchmarks or previous versions of data processing. This helps to make progress in data quality quantifiable.

6. **Feedback integration:**
Collect and integrate feedback from users and stakeholders who work with the processed data. User feedback can provide valuable insights into the practicality of the preparation methods.
7. **Documentation and reporting:**
Document the preparation methods used, test procedures and quality control results. Transparent reporting supports traceability and confidence in data processing.
8. **Continuous improvement:**
Use the findings from quality control and testing to continuously improve the preparation methods. An iterative approach makes it possible to react flexibly to changes in the data or requirements.

Systematic quality control and thorough testing of preparation methods are critical to ensuring the integrity and value of data assets. By applying these best practices, organizations can ensure that their data cleansing and preparation processes deliver reliable, high-quality data that serves as a solid foundation for decision-making and operational processes.



1.3 Data protection aspects in the processing of personal data

The careful handling of personal data, especially in the context of its processing and cleansing, is not only a matter of ethical responsibility, but also a legal necessity. Data protection aspects play a central role in protecting the privacy of individuals and ensuring compliance with legal requirements. The following guidelines and best practices should be considered in order to adequately address data protection aspects in the processing of personal data:

1. **Knowledge of data protection laws:**
A thorough knowledge of the applicable data protection laws and guidelines, such as the European General Data Protection Regulation (GDPR) or similar regulations in other jurisdictions, is essential. These laws define the framework conditions for the collection, processing and storage of personal data.
2. **Data protection through technology design (Privacy by Design):**
Integrate data protection principles as early as the development phase of data processing and cleansing systems. This includes measures for data minimization, encryption and anonymization to ensure the security and protection of data.
3. **Obtaining consent:**
Ensure that valid consent has been obtained from the data subjects for the processing of their data, where this is required. Consent must be specific, informed and voluntary.
4. **Transparency and information obligations:**
Inform data subjects transparently about the type, scope and purpose of data processing and about their rights with regard to their data. This also includes information about data cleansing and processing procedures.
5. **Access and rectification rights:**
Grant data subjects access to their stored personal data and enable the correction of incorrect or outdated information.

6. **Data security:**
Implement appropriate technical and organizational security measures to protect personal data from unauthorized access, loss or damage. This includes security protocols for data cleansing and processing.
7. **Data minimization:**
Limit the processing of personal data to what is necessary for the purpose of processing. Data that is no longer required should be securely deleted or anonymized.
8. **Documentation and verification:**
Maintain thorough documentation of data processing activities, including data cleansing and processing activities, to demonstrate compliance with data protection laws.

Observing these data protection aspects is crucial in order to strengthen people's trust in the processing of their data and to minimize legal risks for the organization. Proactive and responsible handling of personal data forms the basis for data processing processes that comply with data protection regulations.



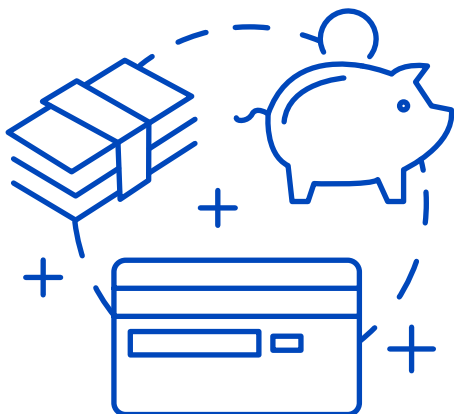
2. Case studies and application examples

2.1 Successful projects for duplicate avoidance through suitable preparation

In a world that is driven by data, the quality of the information that companies work with is crucial to their success. A key problem that affects many companies is the presence of duplicates in their databases. Duplicates can lead to inefficient business processes, inaccurate analysis and ultimately financial loss. Therefore, avoiding duplicates through proper preparation of person search fields is a critical step in data management. Here we highlight two case studies that demonstrate the successful implementation of such projects.

Case study 1: A global financial institution

A leading global financial institution faced the challenge of maintaining the integrity of its customer database. With customers all over the world, the institution's database was filled with millions of entries that contained numerous duplicates. The project team implemented a comprehensive data preparation strategy that included normalization of first and last names, standardization of address formats and normalization of contact data. By using phonetic algorithms to avoid duplicates, the institute was able to significantly reduce the number of duplicates. The result was a clean, efficiently organized database that enabled more precise customer segmentation and more targeted marketing campaigns.



Case study 2: An e-commerce giant

A leading company in the e-commerce sector was faced with the problem that its extensive customer and product database was riddled with duplicates. This was causing problems with customer communication, stock management and ultimately customer satisfaction. The company relied on a combination of software solutions and specific preparation methods for its data. In particular, emphasis was placed on the cleansing and consolidation of address additions and the standardized formatting of telephone numbers and e-mail addresses. The project led to a significant reduction in duplicates and optimized warehousing and customer service processes. In addition, the accuracy of the marketing campaigns was improved, leading to increased conversion rates.

These case studies illustrate that the appropriate preparation of data sets can not only avoid duplicates, but also increase the efficiency of business processes. The keys to success lay in the careful analysis of existing data, the selection of the right tools and methods for data preparation and the continuous monitoring and adaptation of data quality strategies. By focusing on these aspects, companies of all sizes and from all industries can significantly improve the quality of their data assets and thus gain a competitive advantage.



2.2 Analysis of problem cases and their solutions

The analysis of problem cases in dealing with duplicates in databases and the development of effective solutions are crucial steps for companies that want to improve their data quality. In practice, various scenarios arise that present specific challenges and require individual solution strategies. Here we highlight two such problem cases as examples and outline the solutions applied in each case.

Problem case 1: Inconsistent name entries

A common problem that encourages the creation of duplicates is inconsistent name entries. These can be caused by typing errors, different spellings or the use of nicknames. A telecommunications company identified this problem when it realized that customers were listed multiple times in its database under different names, which made customer communication and support more difficult.

Solution approach:

To solve this problem, the company introduced a number of measures, including

- > The implementation of name field normalization to ensure consistent spelling of names.
- > The introduction of an automated check for typos and common variants of names.
- > The use of phonetic algorithms that made it possible to recognize and consolidate similar-sounding names.

Problem case 2: Address variations

Another problem that leads to duplicates is variations in the address details. Different formats, incorrect zip codes and varying spellings of street names can lead to the same address appearing multiple times in a database. An online retailer encountered this problem, which led to difficulties in delivering orders.

Solution approach:

To overcome the challenges associated with address variations, the retailer implemented the following strategies:

- > Standardization of address formats based on national and international standards.
- > Use of address validation tools that check and correct entries in real time.
- > Integration of an address cleansing system that regularly runs through the entire database and identifies and corrects duplicates and incorrect entries.

These solutions demonstrate how the combination of technological tools and clear data management strategies can effectively tackle problems that lead to the creation of duplicates. The decisive factor here is not only the selection of the right tools, but also the continuous maintenance and monitoring of data quality in order to ensure a high level of data integrity in the long term.

List of TOLERANT software products

TOLERANT Software GmbH offers a range of sophisticated tools designed to support companies in various aspects of data management. These products enable improved data quality, compliance with regulatory requirements and more efficient business processes. The following is an overview of TOLERANT's core products and their specific functions.

1. **TOLERANT Match**

Function: This tool is used to identify and merge duplicates within databases. It enables efficient and precise data cleansing, which forms the basis for reliable analyses and business decisions.

2. **TOLERANT Post**

Function: TOLERANT Post ensures the correctness and completeness of address data. It supports the standardization and cleansing of address information, which not only optimizes customer communication but also improves the deliverability of postal items.

3. **TOLERANT Bank**

Function: TOLERANT Bank checks bank data such as account number and sort code and corrects them if necessary. BIC and IBAN are determined automatically.

4. **TOLERANT MPM (Marketing Permission Management)**

Function: TOLERANT MPM manages the consents and authorizations for marketing activities and ensures that all customer communications comply with legal requirements and customer wishes.

5. **TOLERANT PEP**

Function: A specialized tool for identifying politically exposed persons (PEPs) in customer and business partner data. It helps companies to meet compliance requirements such as anti-money laundering (AML) guidelines.

6. **TOLERANT Name**

Function: TOLERANT Name structures name components and generates correct letter salutations so that customers are always addressed correctly.

7. **TOLERANT Sanction**

Function: This tool checks individuals and companies against national and international sanctions lists. It is crucial for compliance and prevents business relationships with sanctioned entities.

8. **TOLERANT Move**

Function: TOLERANT Move determines relocation data of unknown recipients and recognizes deceased persons in the address list.

This product range from TOLERANT covers a broad spectrum of functions, from data cleansing and compliance monitoring to customer authorization management, providing comprehensive support for data management in modern companies.

