

4th TOLERANT Whitepaper

How to find and avoid duplicates in the database

Preparation of people search fields

Chapters 1-2 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. Introduction to the preparation of person search fields

1.1 Importance of search field preparation for duplicate avoidance

In the digital data world, where every byte counts and the purity of information is worth its weight in gold, search field preparation rises like a beacon of order from the stormy sea of data chaos. A key aspect of this shining sentry is its indispensable role in preventing duplicates, those insidious doppelgangers of the data world that not only undermine efficiency but also jeopardize the integrity and reliability of databases. Duplicates - fleeting shadows in the data landscape - are often caused by the smallest inconsistencies in the search fields, be it typos, different naming conventions or varying address formats. These seemingly harmless differences are like traps that lead careless data astray, trapping it in redundancy and confusion. This is where the art of search field preparation comes into play. It is like a careful gardener tilling his field, weeding and ensuring that each plant - or in this case, each piece of data - grows and thrives in its purest form. By standardizing and cleansing names, addresses and other critical search fields, it transforms potential duplicates into uniquely identifiable entities, ready to take their unique place in the database. The preparation of search fields is therefore not just a question of aesthetics or order. It is a strategic necessity that strengthens the foundation of any data

processing operation. By eliminating duplicates, it not only improves data quality, but also optimizes search queries, increases the efficiency of marketing campaigns and strengthens decision-making - a real powerhouse in the engine of data processing. In practice, this means that every data entry is subjected to thorough inspection and processing before it is incorporated into the database. This preparation is a bulwark against the flood of errors and inconsistencies, a protective shield that preserves data integrity and minimizes the risk of misinterpretation. The importance of search field preparation to avoid duplicates can therefore not be overestimated. It is an essential element in the fight against data clutter, a guiding star that points the way to clarity, efficiency and reliability. In a world that is increasingly driven by data, search field preparation is not just a virtue, but an absolute necessity.



1.2 Overview of common challenges and objectives

At the heart of digital transformation, where data is the new oil, companies face the monumental challenge of not only managing their data assets, but turning them into a source of sustainable value creation. The "Overview of common challenges and goals" bullet point highlights the Sisyphean task that lurks behind the apparent calm of data landscapes and outlines the goals that companies are striving to achieve in order to use their data more effectively.

The challenges are varied and complex. First and foremost is duplicate detection and avoidance, a tiresome problem that not only wastes storage space but can also lead to incorrect decisions. The causes are manifold: from inconsistent data entries and different spellings to a lack of standards. Every duplicate is a stumbling block on the way to a clean, efficient database.

Another major obstacle is data up-to-dateness and quality. Outdated information can lead to inefficient marketing strategies, incorrect analyses and ultimately to financial losses. Continuous maintenance and updating of the database is therefore not a luxury, but a necessity.

The integration of data from different sources presents companies with the challenge of creating a coherent, standardized database. Harmonizing data formats and structures is an essential step in breaking down silos and improving collaboration and the flow of information within the company.

The goal is clear: to create a high-quality, reliable database that serves as the basis for business decisions. This requires not only technical solutions, but also a rethink of the corporate culture: data must be seen as a valuable asset whose maintenance and development is at the heart of the corporate strategy.

To overcome these challenges and achieve their goals, companies rely on advanced data cleansing technologies and methods, such as the preparation of search fields, the use of phonetic algorithms for duplicate detection and the implementation of data quality management systems. These tools are essential to ensure the integrity, security and usability of the data.

The "Overview of common challenges and goals" makes it clear that data maintenance is a dynamic, ongoing process that requires flexibility, a willingness to innovate and a clear strategy. In this sense, data maintenance is less of a burden and more of an opportunity: an opportunity to fundamentally improve the performance of companies in the digital age.



2. Basics of data preparation

2.1 Definition and relevance of person search fields

In the digital era, where data is not just a resource but the backbone of business decisions, marketing strategies and customer relationships, people search fields take center stage. These search fields are the navigational stars in the endless galaxy of data records that allow people to be precisely identified and located in databases. The definition and relevance of these search fields are therefore invaluable for any data-driven company.

Personal search fields typically include information such as names, addresses, telephone numbers, email addresses and dates of birth. They are the key attributes that make it possible to uniquely identify individual data records in a variety of applications and contexts. The correct and efficient handling of these search fields is crucial for the quality of data maintenance, the avoidance of duplicates and ultimately for the reliability of the entire database.

The relevance of personal search fields extends across various areas: From customer relationship management (CRM) and targeted marketing campaigns to compliance with data protection regulations. In CRM systems, they enable a personalized customer approach by providing a 360-degree view of the customer. In marketing, they enable the segmentation of target groups and the implementation of precise, customized campaigns. In terms of data protection, they ensure that personal data is handled and protected correctly.

The challenge lies in the complexity and dynamic nature of the data. People move, change their names or contact information, which constantly threatens the timeliness and accuracy of databases. Added to this is the variety of data sources and formats, which makes consistent data maintenance difficult. The preparation of person search fields - i.e. the standardization and cleansing of data - is therefore a critical step in overcoming these challenges.

Investing in the careful management of person search fields pays off. It minimizes the risk of errors, improves customer interaction and satisfaction and increases the efficiency of business processes. At a time when competition is increasingly determined by the quality of data and the ability to use it intelligently, people search fields are more than just data points; they are strategic assets that need to be nurtured and protected.



2.2 Typical structure and content of person search fields

The typical structure and content of people search fields form the foundation for efficient database operations and data analysis in a variety of use cases. These search fields are critical for the identification and management of people records, from customer management systems and marketing databases to personalized services and security checks.

The core elements of the person search fields generally include:

1. **Names:** Probably the most elementary and at the same time most complex search field, which can be divided into first and last names. Names can include prefixes (e.g. titles such as Dr. or Prof.), suffixes (e.g. Jr. or III.), middle names and name affixes. The variety of spellings, cultures and conventions makes the uniform treatment and normalization of this data a challenge.
2. **Addresses:** Another critical search field that is often divided into multiple subfields such as street name, street number, zip code, city, state and country. Addresses are essential for geographic analysis, direct marketing and identity verification. The standardization of address formats and the removal of inconsistencies are essential steps to ensure the usability and accuracy of this data.
3. **Telephone numbers:** These include landline and mobile numbers as well as international dialling codes. Telephone numbers often need to be cleansed of format-related characters and converted into a standardized format to enable effective communication and duplicate recognition.
4. **Telephone numbers:** These include landline and mobile numbers as well as international dialling codes. Telephone numbers often need to be cleansed of format-related characters and converted into a standardized format to enable effective communication and duplicate recognition.
5. **Dates of birth:** Important for age verification, segmentation and personalization. Birth data may be available in different formats and may need to be standardized to ensure consistent and comparable data sets.

These search fields form the basis for the identification and management of individuals in databases. Their structure and content require careful planning and management to ensure the integrity, security and efficiency of the data. Strategies for data cleansing, normalization and validation play a central role in this process, as does consideration of data privacy and security. Effective management of people search fields enables organizations to fully leverage their data assets, provide personalized services and make informed decisions.

2.3 General principles of data cleansing

The general principles of data cleansing are critical to ensuring the quality, reliability and usability of data assets in any organization. These principles serve as a guide for the effective identification and correction of inconsistencies that may affect data integrity. This is not just about "cleaning" data, but about creating a stable and reliable foundation for data-driven decisions.

1. **Consistency:** One of the key aspects of data cleansing is ensuring consistency across different data sets and data sources. This means that the same data should be presented in an identical way, regardless of where or how it was collected. Consistency is particularly important for personal data, such as names and addresses, to avoid duplicates and mismatches.
2. **Accuracy:** Data must not only be consistent, but also accurate. Incorrect data, be it typos, outdated information or incorrect categorizations, must be identified and corrected. The accuracy of the data is crucial for the credibility of analyses and reports.
3. **Completeness:** Incomplete data sets can lead to misinterpretations and inaccurate analyses. Data cleansing therefore also includes filling in missing information to ensure the completeness of the data. This can include the addition of missing address components or the correction of incomplete contact information.
4. **Timeliness:** Data becomes outdated quickly, especially in dynamic environments. A key aspect of data cleansing is therefore updating outdated information to ensure that decisions are made on the basis of the latest available data.
5. **Uniformity:** The standardization of data formats and values is another important principle. This applies, for example, to the formatting of dates and times, telephone numbers and addresses. Standardized data formats facilitate the analysis, comparison and integration of data from different sources.
6. **Deduplication:** Identifying and merging duplicate data records is a key step in data cleansing. This prevents redundancies and ensures that each data record is unique and representative.
7. **Data protection:** Data cleansing must take into account data protection regulations and practices. Personal data must be protected and treated sensitively, and any type of data cleansing should comply with local and international data protection laws.

Applying these general principles of data cleansing enables organizations to effectively manage and leverage their data. By ensuring consistency, accuracy, completeness, timeliness, uniformity and deduplication, while strictly adhering to data protection, organizations can ensure the integrity of their data assets and create the basis for informed decision-making.



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.

