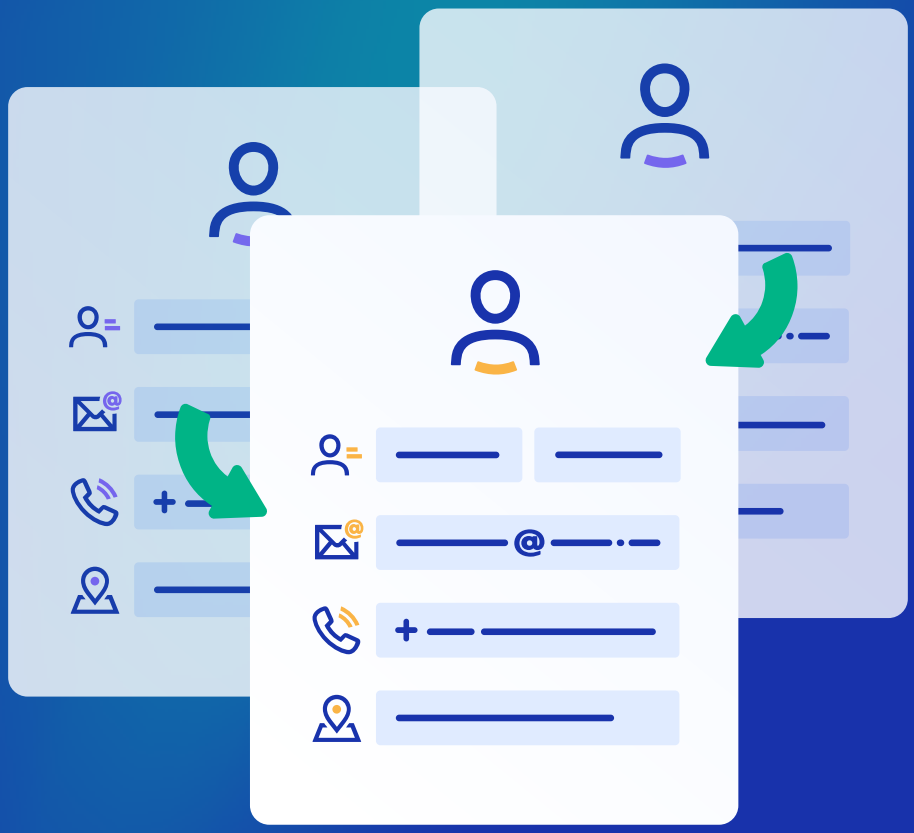
dqe
data quality everywhere

WHITE PAPER

UNIFYING CUSTOMER DATA

A Best Practices Guide to Deduplication and Merging



www.dqe.tech

INTRODUCTION

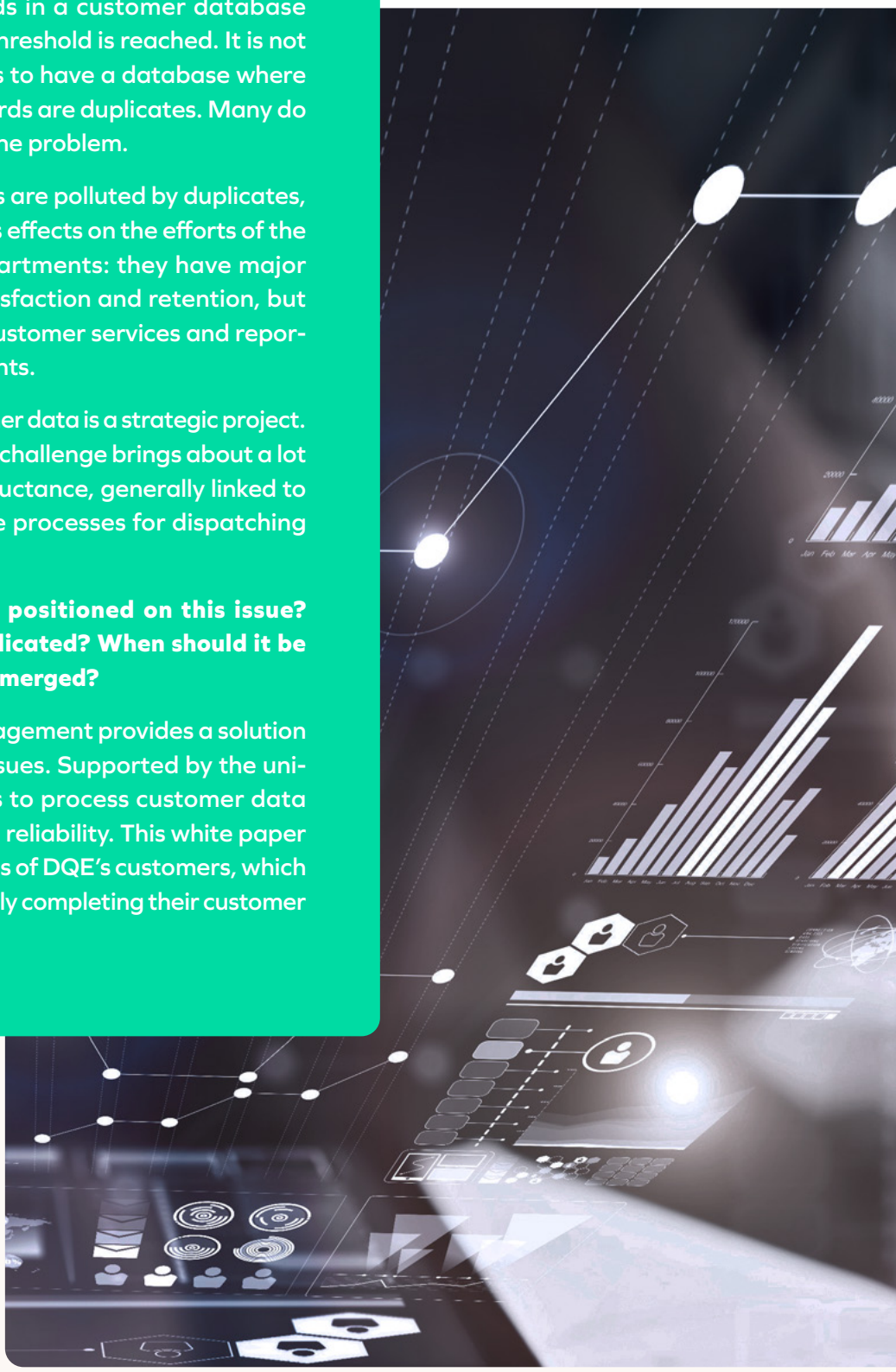
When 10% of the records in a customer database are duplicates, a critical threshold is reached. It is not uncommon for companies to have a database where more than 20% of its records are duplicates. Many do not realize the extent of the problem.

Most customer databases are polluted by duplicates, which can have disastrous effects on the efforts of the company's different departments: they have major impacts on customer satisfaction and retention, but also on the efficiency of customer services and reporting from sales departments.

Therefore, unifying customer data is a strategic project. For many companies, this challenge brings about a lot of questions, and even reluctance, generally linked to the transformation of the processes for dispatching data to the various tools.

Where is your company positioned on this issue? What needs to be deduplicated? When should it be done and what should be merged?

Today, Data Quality Management provides a solution for duplication-related issues. Supported by the unification of data, it serves to process customer data and ensure its quality and reliability. This white paper presents the best practices of DQE's customers, which have led to their successfully completing their customer data unification projects.



1.
Duplicates, a scourge in customer databases

All companies have their own customer database. The quality of this data is a universal issue and handling duplicates is an integral part of Data Quality Management. First of all, it is important to identify the processes that generate duplication.

Where do the duplicates come from?

The major source of duplicates is the compartmentalization of the company's various information systems, which multiply databases - CRM marketing/sales database, customer service, ERP, points of sale, Web portal, etc. This configuration generates duplication in the information system: for each contact point, the collection of customer data is likely to trigger a new customer record. Inexorably, duplicates spread throughout the company. However, each department is aware of the value of quality information.

Other sources of duplication may appear. This is the case of organizational bias, especially when the creation of new customer records is incentivized when trying to develop new business. This practice encourages operational staff to systematically create a duplicate customer record, even if it means using workarounds when an operational tool blocks multiple entries of the same email address. A small marginal change - an extra dot in an email address, a deliberate inversion of letters - allows you to «force» the system, which will identify the record as different.

The consequences of duplication can be disastrous:

Customer service, call centers and points of sale

Duplication puts pressure on those in contact with customers. During a phone call, several customer records appear on the screen for the same customer, which one should the representative choose? Which one is the right one? Which form should be filled in?

It is impossible to have a 360° view of the customer, of their interactions with the company - orders, calls to customer service, visits to the office or point of sale, previous complaints, etc. How can you give quick and relevant answers, the main mission of customer service? This duplication becomes a major obstacle to customer service efficiency.

Marketing and CRM departments

These departments can hardly conduct campaigns with a correct segmentation of their customers if customer data is scattered over different accounts.

Moreover, a customer database full of duplicates can quickly lead to repeated mailings to the same people, or inappropriate messages, for example an incentive for a first purchase sent to a regular customer. When a company maintains a premium brand image, these approximations go against the message.

Sales force figures

They suffer due to numbers that rely on duplicate customer data. For example, if the purchase data for a customer is scattered over several files, it is impossible to know the revenue generated by the person and to identify them as a high-potential customer.

Similarly, the opportunity conversion rate goes down when calculated on a customer database that has a number of duplicates that is 10% or higher. Reporting, visibility, numbers: everything is distorted.

Is your company affected?

You may have already taken steps to reduce the number of duplicates in your customer database, but is the result adequate? Ask your field staff: they are in the best position to confirm the presence of duplicates that hinder their operational efficiency.

First and foremost, you need to understand the extent of the problem in your company. To find out how many duplicates there are in your customer database, ask a specialist to audit a sample of your database. The results may be hard to swallow, but they justify taking action.

2. Setting the stage

A project to unify customer data concerns the entire company. It is therefore necessary that all services be coordinated in the effort. In addition to good project management and change management practices, it is important to work more specifically on data unification.

Best practices

Accepting opening up

The data culture in many companies is still based on a reluctance to share data internally. For each service, its data is its own private domain. When several systems coexist, each one with its own adatabase, it is impossible to deduplicate, merge and create a single customer repository - the search for the Golden Record becomes an uphill climb.

Achieving the Golden Record requires deduplicating each existing database, then reconciling them into a unified customer repository where validated customer data is merged.

Define a sponsor

The sponsor is a key factor in the success of the project. Depending on the company, this role may be entrusted to the marketing director, the customer service director, the financial director, etc. This sponsor must have “free rein” and ensure that each department cooperates in making its data available. In return, the sponsor assures that the resulting Golden Record will be accessible to all.

Adjust expectations in the project

In terms of deduplication, perfection does not exist. To put it another way, aiming to identify all the duplicates in a database is not a realistic goal.

The identification of duplicates is a question of compromise between fault tolerance on the one hand, and the percentage of proximity between records on the other. The support of an expert is imperative here to know the elements on which your company should focus its efforts.

3. Identifying duplicates

What does your company consider a duplicate?

This is the first question that needs to be asked. There is no universal answer, because everything depends on the company's business. The first step is to identify the most relevant data fields for comparison. These fields will then be used as reconciliation keys to identify duplicates.

Best practices

Define at least three duplicate identification fields

Define at least 3 duplicate identification fields to achieve effective deduplication. If the fields such as last name, first name, and email address are part of the commonly used elements, companies can also define others that are just as important, depending on their business: social security number in the health sector, vehicle registration number in the automotive sector, registration number and company legal information in BtoB databases, etc. The fields selected in this way must have priority in order to identify a single customer.

To identify as many duplicates as possible, it is best to screen the database for each reconciliation key that is defined. In other words, it is a matter of using three consecutive filters to tighten the focus for identifying duplicates.

Introduce the notion of "fuzziness"

There is no point in looking for duplicates based on criteria of exact similarity between records. Only a few duplicates will stand out at most.

Hence the notion of fuzziness, i.e., the exploration of approximations in data entry that cause the multiple records of a single customer to be identified as different by business systems such as CRMs or ERPs - first name in the form of an initial and then entered as a full name on another occasion, landline and mobile numbers reversed between different records, typo in a name, addresses of different residences, etc. Fuzziness takes many forms! This concept has to be introduced in order to detect more duplicates, which also requires the support of a specialist.

4. Merging

Two questions guide the logic used for merging: what needs to be merged and when should it be merged?

What needs to be merged?

Which of the data elements included in the duplicates of the same customer have priority and will "overwrite" the others in the unified parent record?

Best practices

Keep the oldest duplicate

Keep the oldest duplicate as the base record, as it is usually the most complete.

For the other duplicate records, keep only the most recent data, including email, address, telephone, and related items such as orders, service calls or appointments, for example).

Take the source of the data into account

Data entered by the customer on their online account will be much more reliable than data from a marketing listing.

Discard "noise" words

Discard "noise" words which, unlike the reconciliation keys, are of no use for merging. For example, in a BtoB database of transport companies, many will include the terms "Transportation" or "Logistics" in their company name. The company name field must therefore be removed from consideration, since it cannot be a relevant criterion for merging customer records.

When should the merge occur?

Should we merge in real time, or in batch mode in regular processing on the database where duplicates may have been created?

Best practices

Adapting to the configuration of the information system

The more a company has decompartmentalized its IS and interconnected its various solutions, the better it can opt for the most qualitative solution: merging customer data in real time at the moment it is collected.

On the other hand, in companies where the various solutions are compartmentalized, batch mode is more appropriate, provided that the deduplication/unification solution is interfaced with the various systems. In this case, regular processing of the database makes it possible to clean up its duplicates according to a chosen frequency.

And when not to merge!

Some use cases require maintaining multiple records for a single customer. This is the case of a multi-brand group whose customers frequent its various stores - a record must be kept for each one. Similarly, when customers have multiple residences - in the luxury sector in particular - each one is important information to have. However, the format of the customer records in the CRM tool may not have provided enough address fields to be able to record them all. Here again, different records have to be kept for the same customer.

The trick is to tolerate duplicates without creating confusion for users: the Golden ID, a unique identifier for each customer that is common to all of that customer's records.

Dare to merge!

Some companies know they need to unify their customer data, but hesitate at the merger stage. Their fear: the risk of losing customer data - in other words, merging two distinct customers into one record. Merging is the final step in the customer data unification project.

Your company can count on several safety nets in the art of merging customer data:

Manual merge

For companies that prefer not to automate everything and leave decision-making to departments, it is always possible to maintain the option of a manual merge. Thus, once duplicates are identified, the merger depends on human intervention, which some companies find more reassuring.

The score calculation

The higher the score, the more the score indicates duplicates are present. The score calculation is carried out according to the different fields used as reconciliation keys, and involves defining the threshold at which a company considers that a duplicate is present. This indicator requires good control of the weight of the reconciliation keys used. For example, a high score may hide a single discrepancy that automatically disqualifies the presence of duplicates if it relates to a data field that is of priority importance in your business. Once again, the support of your expert is essential to defining the appropriate rules.

Consistency check

Once the duplicates have been identified, the consistency check helps to define which data elements to keep in the parent record, and which to delete because they are incorrect.

An example: the elements preceding the @ in an email address checked as valid are often the customer's first and last name. They offer a relevant comparison criterion to check the exact spelling of the last name and first name, when these two pieces of information present differing spellings between duplicates. The consistency check with the email address confirms the exact spelling of the first and last name to be kept on the parent record.

The percentage of proximity

The percentage of proximity between the different records defines the similarity threshold at which we decide to merge the duplicates.

Unifying customer data: never without an expert!

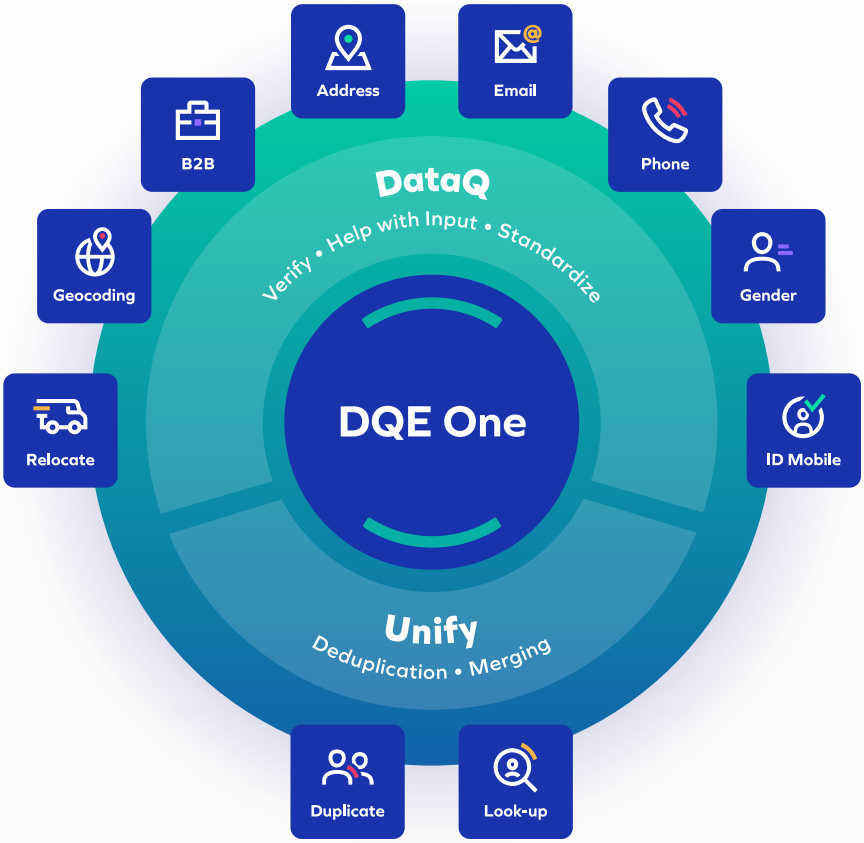
Believing that you can develop a solution to unify customer data on your own is a trap you should avoid! These are usually long-term projects with results that are very uncertain or even wrong.

Our experience with several companies confirms that

All those who have tried to deduplicate and merge their customer data on their own have lost huge amounts of time and money in development, with results that fall far short of the requirements. It is not uncommon to see processing times of several weeks with in-house solutions, when a few hours are sufficient with a solution developed by a specialist. Moreover, the identification of duplicates leaves much to be desired when a solution is too generic, or when it does not use reconciliation keys that are granular enough to find a maximum number of duplicates in a database. Finally, there is the case of companies that stop the process when it comes time to merge due to lack of control and fear of losing customer data.

To bring your customer data unification project to a successful conclusion, enlist the help of an expert like DQE

DQE's Data Quality offer is based on 15 years of experience, proven technologies that are constantly being optimized, and a thorough knowledge of best practices, points where attention is necessary, and pitfalls to avoid. This way, you can be sure to unify your data knowing all the facts, to improve your efficiency, and ultimately, your results.



About DQE

Founded in 2008 in France, DQE Software (DQE) is the vendor of DQE One, a modular Data Quality Management solution.

Our solutions are specialized in BtoC and BtoB customer contact and identity data. DQE One controls the quality of data on information such as name, first name, email, telephone, address and company information. DQE One has two product lines: DataQ, a contact data quality solution, and Unify, a deduplication solution.

Across all industries, DQE helps organizations ensure that their customer data is reliable, accessible to everyone, every time. DQE One stands out as the most powerful data quality solution on the market, capable of managing databases of several tens of millions of contacts. Each year, our processing represents more than 3 billion requests with an average response time of 150 milliseconds.

DQE One relies on powerful engines and algorithms that benefit from over 10 years of experience. The solution interfaces with more than 240 international databases for the postal address part. Natively, DQE One offers connectors with solutions such as Salesforce, Microsoft/Dynamics, Cegid, Magento, etc. Our algorithms work on all our clients' clouds: Heroku, Azure, AWS or GCP. DQE One is used by a wide range of business lines (CRM teams, Chief Data Officers, Chief Digital Officers, CIOs, Customer Relationship Managers) at more than 400 customers. Among them are several Fortune 250 companies and renowned brands such as Groupama, EDF, BUT or Belambra.

For more information, visit www.dqe.tech and follow us on

ISBN : 9782494187023