

# High Precision Matching at the heart of Customer Data Integration

The quality of your CDI Solution is only as powerful as the quality of your  
matching engine

## White paper

Release date: September 2008

Version number: 1.0

Author: Emile van de Kloek

Co-author: Holger Wandt

Copyright © 2008 Human Inference Enterprise B.V., Utrechtseweg 310,  
6812 AR, Arnhem, The Netherlands. All rights reserved.

The text and materials in this whitepaper are protected by intellectual  
property rights and remain property of Human Inference.



Human Inference

## Table of contents

The high costs of low quality customer data spread across multiple systems	3
The world upside down – Data mismanagement throughout your entire company	4
Bottlenecks causing a low quality and incomplete customer view	4
How to create a single customer view?	5
The heart of any architectural Customer Data Integration solution	6
The truth about high-precision matching methods	7
Combining deterministic and probabilistic matching – a hybrid approach	8
Interpretation based on natural language processing	8
Convert high cost into high value with High Precision Matching at the heart of your CDI solution	9

## The high costs of low quality customer data spread across multiple systems

There are many examples (and you can probably think of some from your own experience) where poor, fragmented or defective customer data cost time and money or cause severe frustration or even bodily injury.

Here are just a few examples from the author's own experience that (painfully) illustrate this data integrity and integration issue:

- Life insurance contribution deducted from the wrong bank account even after notifying the insurance company officially, four times, with a request for change.
- A support engineer driving to the right customer but to the wrong address because the address information was only updated in the CRM system and not in the separate support system.
- Getting full access to an online newspaper because the wrong email address is attached to the right customer. This customer receives the invoice but doesn't have access to his online subscription.

These examples show the impact on just one individual. But there are many public examples in which large groups of people are involved and / or the impact and costs are huge:

- 730.000 Dutch citizens were duped because the Dutch tax collectors office lost their tax filings. The combination of an old tax return system and a new system for digital signatures caused this big problem.
- A US B2 bomber hit the Chinese embassy in Belgrade (1999). Due to an outdated map the right address was applied to the wrong building.



These lists of examples only reflect the tip of the iceberg. Most data integrity and integration issues are hidden in day-to-day work or they are being perceived as a fact of life. What about your experiences? Ask yourself the following questions:

- How many separate databases and independent applications in your organization contain customer data which can not be updated centrally?
- How many proprietary and private customer databases in your enterprise reside on laptops and computers of employees (MS Excel, MS Outlook, PC Databases, etc), including information that is not integrated with and synchronized to your central customer database like your CRM system?
- How many times are you, or even worse, your customers, frustrated because of time consuming actions and rework and the negative impact involved?

Data audits reveal that invalid data values in customer databases average around 25 to 30 percent. The direct costs involved by these invalid data values are usually tangible. These costs are all related to the actions that involve not doing jobs right the first time; i.e. sending mailings or magazines to the wrong addresses or manually assembling customer data across disintegrated databases. However, the most significant real costs are much higher. What happens if a customer is so frustrated that he or she decides to switch to your competitor? Then we are talking about lost customer lifetime value<sup>1</sup>. Lost or missed customer lifetime value as the result of poor customer data quality and a lack of a central view on the customer can be significantly greater than the money wasted on duplicate and wrong address mailings. For example, wasted mailing costs of Euro 5.000 may actually result in millions of euros in lost customer lifetime value.

### **The world upside down – Data mismanagement throughout your entire company**

Think of all the resources in your company. Almost all resources are consumable or can only be used one at a time. Money can only be spent once, employees can only perform one task at a time, raw materials within a production process can only be used once, and your meeting room can be used for many different purposes but can only be used for one purpose at any given time.

Data resources like your customer data, however, are resources that are completely reusable. In fact, these are the only resources where high redundancy is accepted as a legitimate cost of doing business.

*The question arises as to why organizations have come to accept this wasteful and costly approach?*

This redundancy sounds absurd. Imagine yourself hiring multiple employees to perform exactly the same job or paying a single invoice multiple times or leasing multiple company cars while you only need one. However, these examples are actually equivalent to your customer data redundancy in which you use multiple databases and applications capturing the same customer data by different information producers.

Customer data quality and integration problems impact virtually every area of the value chain of your business. From primary activities like inbound- and outbound logistics, marketing, sales and operations to supporting activities like procurement and human resources. All the time and money your business spends on, for instance, the following activities are wasted and not available for adding value to your company:

- Adding the same customer information manually in multiple databases
- Correcting inaccurate data
- Building workarounds for customer data problems
- Searching for missing data
- Manually enriching customer data in one system or the same customer data in multiple systems
- Assembling customer data across disintegrated databases
- Resolving customer data related complaints

The question arises as to why organizations have come to accept this wasteful and costly approach?

<sup>1</sup>Customer lifetime value is the net present value of the future cash flows attributed to the total customer lifetime relationship

## **Bottlenecks causing a low quality and incomplete customer view**

Of course people make mistakes when entering customer data into databases, causing this data to be inaccurate and incomplete. This will always be the case. In addition, many companies rely on an IT-infrastructure based on multiple databases and applications.

This silo structure is very common, simply because companies need to facilitate the specific business processes within the primary and supporting activities in the total value chain of the organization. Companies typically depend on several ERP, CRM, Customer Service, Invoicing and other front- and back office systems. Unfortunately this collection of separate databases and applications are often not or only poorly integrated. This results in customer data being scattered across the enterprise. Data quality software can solve many issues within single silos but it doesn't provide an accurate and complete view of every individual customer throughout multiple silos.

So it is not that organizations just accept this wasteful and expensive upside-down approach. Very often the pool of multiple separate databases and applications has grown through time due to specific business needs. While the philosophy of CRM was to aim at a single system containing all customer information and processes related to customers, it turned out that this approach failed in many organizations. To make matters even worse, if you think you finally have the right solution in place, this can be completely disturbed by a merger or acquisition.

## **How to create a single customer view?**

It is obvious that enterprises need a complete or single customer view to avoid the inextricable consequences; i.e. the high costs involved in not providing this single point of truth to all primary and supporting activities in the company. In addition, these costs are caused by poor quality customer data divided over disparate and separate databases and applications in the organization. It was concluded that CRM was supposed to provide us with the answer to this problem, but it didn't. So the question is what to do about it?

*CDI enables companies to achieve a single customer view and to simultaneously improve the agility of existing applications*

CRM as the tool for creating a single customer view failed because of the single system philosophy. The antithesis to a single system philosophy is the multiple systems philosophy. Within CDI (Customer Data Integration) we accept the fact that almost every company has to face the legacy of multiple disintegrated systems supporting different business processes. A CDI solution also allows future changes, driven by business needs, within your IT infrastructure; i.e. a merger or acquisition results in the need to add a totally different IT infrastructure to your existing one. Instead of replacing the existing and / or additional infrastructure and systems, CDI provides a CDI hub as a central repository for customer data which is connected to the existing infrastructure of multiple systems. Also future additional systems can be connected to the CDI hub. The customer data in the central hub is synchronized with the customer data in the original source systems, in batch mode or real-time.

So CDI enables companies to achieve a single customer view and to simultaneously improve the agility of existing applications. But what is the secret of a high quality single customer view?

## The heart of any architectural Customer Data Integration solution

Customer Data Integration, also known as Master Data Management for Customer Data, provides companies with a sustainable answer to the high costs caused by their inaccurate disintegrated database environment. According to Gartner's definition<sup>2</sup>, CDI is much more than just technology:

*"A combination of technology, processes and services to deliver an accurate, timely and complete view of the customer across multiple channels, lines of business, departments and divisions drawing customer data from multiple sources and systems"*

Within this paragraph we only focus on the technology part of CDI and more specifically on the kernel of the technology: the matching engine. The matching engine takes care of the mapping and grouping of customers within and across multiple heterogeneous databases. When two or more records spread across this pool of multiple databases belong to the same customer, the matching engine should be capable of recognizing this and assigning those records to the same group. Without this matching technology it's not possible to deliver an automated single customer view to your company. With this matching technology it's possible to create a central repository for customer data containing the so-called "Golden Record" and references to the original source systems.

Based on the purpose of the single customer view different architectural CDI designs are necessary which can differ from company to company. Gartner defined four architectural CDI designs<sup>3</sup>, see table 1: "Overview architectural CDI styles based on Gartner".

	<b>Consolidation style</b>	<b>Registry style</b>	<b>Coexistence style</b>	<b>Transaction style</b>
<b>Storage Golden Record</b> (consolidated view of master data)	X (Not up-to-date)	- (only cross reference Index to source records)	X (Not guaranteed up-to-date)	X (up-to-data)
<b>Published view</b>	-	X (Dynamic)	X	X
<b>Central Authoring</b>	- (Authoring remains distributed)	- (Authoring remains distributed)	- (Authoring remains distributed)	X
<b>Suitability</b>	<b>For reporting, analysis and central reporting</b>	<b>Mainly for real-time central reference</b>	<b>For harmonization across databases and for central reference</b>	<b>Act as system of record to support transactional activity</b>

← Analytical Focus
Operational Focus →

Table 1: "Overview architectural CDI styles based on Gartner"

<sup>2</sup>"Creating the Single Customer View with Customer Data Integration", John Radcliff, Gartner, Inc. January 9, 2006.

<sup>3</sup>"How to Choose the Right Architectural Style for Master Data Management", John Radcliff, Andrew White, David Newman, Gartner, Inc. September 28, 2006.

Some companies just need a central and unique view on their customers for reporting and analysis (Consolidation style). Other companies are not allowed to store customer information within a central repository and only use the cross reference index to create a dynamic view of the customer (Registry Style). Most organizations use CDI for harmonization across databases and for central reference (Coexistence style). Finally within the ultimate solution CDI provides central authoring and guaranteed up-to-date customer information to support transactional activity (Transaction style). Despite these separate styles the implementation of a mix of styles to serve different purposes is also possible.

Regardless of the purpose of the single customer view and the associated architectural CDI design: at the heart of every CDI solution is the matching engine. Within the matching engine the secret of a high quality single customer view is concealed. The quality of the single customer view and the resulting business values depends on the power of the matching technology. If, for instance, the matching engine is not capable of recognizing that **BMW** and **Bayerische Motorenwerke** is actually the same customer you will still accrue unnecessary costs or miss value creating opportunities (opportunity costs) within your company. Because the quality of the matching engine is so important we distinguish “standard matching” from “High Precision Matching”. In this whitepaper we want to emphasize the need for High Precision Matching at the heart of any Customer Data Integration solution. In the following paragraphs we explain the ingredients and requirements of High Precision Matching technology.

## The truth about high-precision matching methods

There are many theories on matching, but in general there are two methods that prevail in customer data integration systems: the deterministic and the probabilistic approach. Both approaches to matching have advantages and disadvantages, but they also have one thing in common. The higher the level of domain-specific and statistical knowledge, the better the assessment of the degree of similarity between database records.

- Deterministic matching uses, among others, country- and subject-specific knowledge, linguistic rules, such as phonetic conversion and comparison, business rules and algorithms, such as letter transposition or contextual acronym resolving to determine the degree of similarity between database records.

Example: The match between **EVO AG** and **Energieversorgungsgesellschaft Offenbach** is in part determined through contextual acronym resolving and subject-specific knowledge on legal forms and compound words in the German language: AG in German is the standard abbreviation for the legal form *Aktiengesellschaft*, of which the word *gesellschaft* is frequently used in compound words – a very common process in the German language.

NB: This example is based on extensive domain-specific knowledge, which enhances the quality of the particular deterministic matching method.

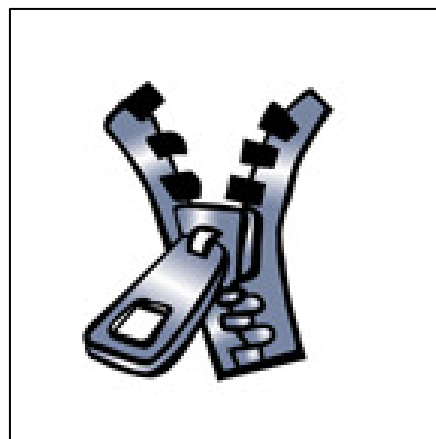
- Probabilistic matching uses statistical and mathematical algorithms, fuzzy logic and contextual frequency rules to assign the degree of similarity between database records. In this, patterns with regard to fault-tolerance play an important role (the matching method is able to take into account that humans make specific errors). Probabilistic matching methods usually assign the probability of a match in a percentage.

Example: The word *London* can have different significations. It can, for example, be a surname or it can have a geographical meaning, being the capital of England. However, the probability of London being a surname is much lower than it being a geographical indication. Compare **Jack London Ltd.**, **Thompson London Ltd.** and **The London Consulting Group Ltd.**

## Combining deterministic and probabilistic matching – a hybrid approach

Within the CDI market space, there are a lot of opinions on the difference in accuracy and performance of the aforementioned matching methods. This is, however, an insignificant discussion.

Real high quality, high precision matching engines that deliver the required results must make use of deterministic and probabilistic matching. The reason for this is actually quite simple: the better the matching engine is able to determine *what is what* in a particular context, the better the probability calculation of a certain match or a certain non-match. This is, in essence, the same as humans do. We determine what we know and consequently use contextual probability and pattern recognition to assign significations to the words we come across: **John Edward Smith** is the combination of a very frequent first name, a common middle name and a highly frequent surname. This knowledge is a clear advantage when interpreting the string **John Edward Agandong**, where the surname has a very low absolute frequency. Using contextual probability however, we are able to assign the signification surname to this particular word.



When comparing the numerical strings **050512** and **12052005**, it is beneficial to know that there are different types of date notation, since this definitely increases the similarity probability of the two strings. A straightforward mathematical comparison would generate a lower probability rate.

## Interpretation based on natural language processing

In order to correctly interpret business data, the interpretation engine must analyze the data in a “human fashion”. This is called natural language processing. The analysis consists of:

- Tokenizing: the data is tokenized into separate words
- Characterizing: each word is assigned attributes, such as pronouncibility and prosodic structure
- Classification: each word is assigned a categorical classification (i.e. syntactic class = proper name)
- Grouping: all possible groupings of categories are generated (i.e. a + b yields c, where a, b and c are categories).
- Path selection: the optimal path through the (many) groupings is selected and the signification is assigned.

Automated interpretation of business data is not a simple task. It must, for instance, deal with the fact that many words have more than one meaning and that data in different data sources is often non-standardized, incomplete, incorrect or otherwise mutilated.

In CDI, matching over a variety of data sources is common practice. Combining deterministic and probabilistic matching will yield more precise matching, with less false positives (mismatches) and less false negatives (missed matches). Probabilistic matching often uses weighting schemes that consider the frequency of information to calculate a score and/or ranking. The more common a particular data element is, the lighter the weight that should be used in a comparison. That is a sound and robust approach. However, assigning weighting factors on data that have been interpreted **and** enhanced with statistical information will increase the matching results to a high precision level.

*It is not a question of choosing either a deterministic or a probabilistic matching method. It is a question of choosing a hybrid method that will satisfy your CDI requirements in the best possible way.*

It is not a question of choosing either a deterministic or a probabilistic matching method. It is a question of choosing a hybrid method that will satisfy your CDI requirements in the best possible way. Of course, such a method must include natural language processing capabilities and domain-specific knowledge repositories.

### **Convert high cost into high value with High Precision Matching at the heart of your CDI solution**

High Precision Matching technology at the heart of any architectural CDI solution, as delivered by Human Inference's product Hlquality Identify, makes it possible to convert high costs into high value.

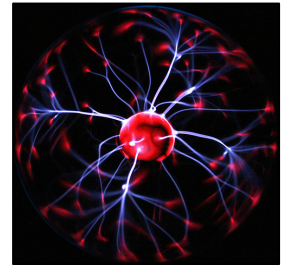
First, all costs related to the processes and actions that involve not doing jobs right the first time as the result of the lack of a (high quality) single customer view can be reduced to an absolute minimum; i.e. assembling customer data across disintegrated databases manually and resolving customer data related complaints. Second and even more important, the possible threat of lost customer lifetime value as the result of dissatisfied and frustrated customers switching to competitors can be abolished and result in possible savings of millions of euros.

Next to these cost saving aspects we have to consider the value of opportunities created. Time and money spent on unnecessary processes and activities can be used to add value to your company. After delivering the single customer view by a central repository of customer data it is possible to create additional value adding activities like:

- Reliable reports and analyses can be made to support reliable decision making
- The right number of mailings or catalogues can be sent to the right addresses
- Portals like, for instance, a call center portal with a complete view on your customers can be created for improved customer experience, cross- and up selling and improved risk management
- Requests for changes like address information or account numbers will be processed right the first time
- Credit risks are available to employees preventing high risk accounts receivable.

It is also possible to adapt your existing source systems to interact real-time with the central repository to deliver up-to-date and complete customer information to the channel, lines of business, departments and divisions supported by these original source systems. The creation and monitoring of high customer experience, with the help of a single customer view, resulting in satisfied and loyal customers is the foundation for even more value creating activities:

- It helps in the process of market segmentation and targeting.
- The right knowledge and a complete view of your customers serve as an important source for product and services development.
- The combination of targeting the right market and right customers with the right products and services is the basis for a sustainable competitive advantage



Whatever architectural CDI style you choose, the High Precision Matching engine is at the heart of the solution. So if you choose to start with the consolidation style for simple customer data related reports providing early value to your company, you're still able to transform that CDI solution to a higher level architecture like the co-existence style or transactional style or even to a multi domain Master Data Management solution, using the same high precision matching engine.

*The quality of your CDI solution is only as powerful as the quality of your matching engine.*

The quality of your CDI solution is as powerful as the quality of your matching engine. If your matching engine has a quality level of only 80% your CDI Solution can only deliver a maximum quality level of 80%. So an investment in a high precision matching engine at the heart of your customer data integration solution can be seen as a very important and sustainable resource for a high quality single customer view. This high quality customer view will help you to reduce costs and to generate value.