



Initiate™

► WHITE PAPER

Customer data integration
and accurate data matching



► WHITE PAPER

Customer data integration and accurate data matching: achieving a 360° customer view

Executive Summary

Why do you receive calls from your existing long-distance telephone company asking you to switch to their service or receive promotions from your bank for services you already use? The harsh reality is that for many companies, data integrity is so poor that they have no idea who a significant number of their customers are. The underlying cause for this confusion is the variation in customer identities, such as discrepancies in name, address, numerical identifiers and other customer unique attributes. Ultimately, this lack of data integrity translates into lower customer satisfaction, wasted resources and compliance risks.

To make matters worse, most organizations have multiple information systems or databases that are either poorly integrated, or in many cases, not integrated at all. This fragmentation creates a major barrier to ever achieving a consolidated view of the each customer – the foundation for successful Customer Relationship Management (CRM), Business Intelligence (BI), data warehousing, and other customer-centric initiatives.

This white paper describes how, with a customer data integration solution that leverages an accurate matching process, you can link the data in disparate information systems to gain a holistic 360° view of each of your customers on demand, while increasing data integrity in each data source. It focuses on how Initiate Systems' proven Initiate Identity Hub™ software achieves the highest levels of accurate data integration on-demand, regardless of the volume of data. To date, Initiate's software and matching process have been used to successfully match billions of records. The paper concludes with the critical criteria for a customer data integration solution and gives details around candidate selection, comparison techniques, probabilistic scoring, and dynamic threshold capabilities.

Introduction

Has your existing long-distance telephone company called you at dinner to ask you to switch to their service? Does your bank send you multiple mailings, including ones for services you already use? Is there a catalog company you buy from frequently that sends you multiple catalogs? Have you ever taken your child to the emergency room at the hospital, only to have the registration clerk ask you for the same basic information you gave on your last visit?

Unfortunately, these situations occur for the same underlying reason – the organization does not have its act together in terms of effectively integrating its customer data. Even worse, these examples are the norm, not the exception. In 2002, The Data Warehousing Institute (TDWI), a for-profit research center for the data warehousing and business intelligence industries, surveyed 647 respondents who quantified costs saved by corrected errors in customer files. It found that the current level of data quality costs U.S. businesses \$600 billion per year in printing, mailing costs and staff overhead alone. Equally shocking is the finding by industry analyst firm Gartner: Through 2006, more than 50% of CRM initiatives will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues (0.7 probability).¹

To optimize business operations, capitalize on opportunities and increase customer service – as well as maximize the return on investment from Customer Relationship Management (CRM), Business Intelligence (BI), data warehousing, and other customer-centric initiatives – organizations typically face two key issues: (1) integrating disparate information systems to aggregate data and gain a holistic 360° view of each customer, and (2) increasing the level of data integrity. These two challenges are closely related as data integrity issues typically manifest themselves when organizations attempt to aggregate data across the enterprise in an attempt to become more customer-centric.

Roadblocks to a 360° customer view

Many organizations have multiple information systems or databases that are not integrated or are poorly integrated. According to research firm IDC, organizations have on average 49 applications and 14 databases that need to be integrated, and typically no more than 20 percent of customer data residing in one location.

This silo structure is not surprising as information systems often support a specific business process (marketing, sales, customer service) or a specific line of business (life insurance, health insurance, automobile insurance). The challenge here is that information on the same customer may be included in each system with minor or even major variations in the data. With no integration and "cross talk" between systems and no data aggregation, it is almost impossible to build and access a complete and accurate profile of each individual customer.

In addition, organizational issues often impede standardization and change. As new information systems are developed to address specific business requirements, they often result in the creation of new data silos. Frequently, for reasons ranging from architectural incompatibilities to data "turf wars," these silos are not integrated with other existing data stores and applications.

Coming to grips with data integrity

This silo structure that is so common also affects the many aspects of data integrity (accuracy, completeness, timeliness, relevance, etc.). This is not surprising as most organizations build and deploy IT systems with a level of data integrity just high enough to accomplish a fairly basic, but specific task – customer billing, for example. When the organization attempts to use data stored in these systems to support a business process with more demanding data accuracy or integration requirements, lack of data integrity becomes an issue.

At this point an organization has two options: (1) it can either undertake an enterprise-wide data quality improvement effort, which is extremely complex and time consuming, in part because it is extremely hard to gauge "how much standardization and quality is enough," or (2) it can employ a data integration solution which uses an accurate matching process. The big advantage of the second option is that an organization can accurately aggregate data without having to undertake a major data improvement effort. Moreover, if it is done well, the process of aggregation itself improves data integrity because it yields a complete information set about each customer.

Matching accuracy is defined by two key criteria: (1) collating a complete set of customer information even if, for example, the individual is known as Thomas Jones in one system, Tom Jones in another, and T. S. Jones in a third, and (2) avoiding mixing information from two distinct customers that live in the same town, for example Thomas Sterns Jones and Thomas Steven Jones. For vendors in the customer data integration (CDI) business, matching accuracy is defined by the performance achieved against both types of potential matching errors.

Of course, accuracy does not exist in a vacuum. Accuracy requirements should always be derived from business requirements. Not matching two records which should be matched has a much different cost in creating a mailing list (where it is essentially the cost of an additional mailing) than in a healthcare delivery organization where it may mean that a record of a patient's drug allergies are unavailable to the attending physician. Similarly, linking two records incorrectly in the mailing list example means that a potential customer does not receive your promotional material, while it may mean that a physician has incorrect information in developing a suitable course of treatment.

Initiate Systems offers solutions to both data aggregation and accurate data matching with its state-of-the-art Initiate Identity Hub™ software, supported by advanced proprietary algorithms and sophisticated matching techniques.

Unifying multiple customer records: Initiate Identity Hub™ software

Initiate Identity Hub™ software serves as an essential foundation on which to launch successful CRM, BI, Enterprise Resource Planning (ERP), and other IT-based business initiatives. While these systems generate and store massive amounts of data, too often they do not share important data. Initiate Identity Hub™ software breaks down barriers between systems and provides a critical link between data sources and operational applications.

Specifically designed to consolidate fragmented customer identities into a linkage set, this customer data hub provides a true 360° customer view by accurately identifying and instantly linking the records for each customer. As a result, it is possible to provide the clean, complete personal profiles required for more cost-effective and responsible interactions with customers, patients or partners.

Initiate Identity Hub™ software accommodates data in real-time or batch files from individual records, internal data sources, business partners or other third-party sources. Data enters Initiate Identity Hub™ software in its native format. From this, new data is derived that enables fast, accurate and efficient linking. Proprietary algorithms compare records and produce links and scores that indicate which records likely represent the same entity and how strong that likelihood is. The algorithms use a combination of techniques – which are outlined in the next section – to improve accuracy far beyond other types of searching applications.

Initiate Identity Hub™ software posts results to the existing applications with which it is linked. Depending on which Initiate application is installed, Initiate Identity Hub™ software can place any exceptions or unresolved cases into a work queue for review and resolution, or it can initiate other business rules, messages and alerts.

Improving data integrity through accurate data matching

Developing a highly accurate matching process requires understanding of the types of errors and their root cause. Formally these errors are:

- ▶ False negatives – where two records which relate to the same member are not linked during the matching process, and
- ▶ False positives – where two records that do not relate to the same member are linked during the matching process.

In the language of statistical decision theory, matching is a hypothesis test. Two hypotheses are being tested:

H_0 : The records do not refer to the same member

H_1 : The records do represent the same member

In this terminology, false positives and false negatives are the Type I and Type II errors, respectively.

DEVELOPING A HIGHLY ACCURATE MATCHING PROCESS REQUIRES UNDERSTANDING ERROR TYPES AND THEIR ROOT CAUSE.

		Matching decision	
		Match	Don't match
Truth	Same member	Correct decision	False negative
	Different members	False positive	Correct decision

Matching systems strive to reduce the total number of incorrect decisions. This requires an understanding of the error source.

False negatives arise from variation in recording demographic information. Typical types of variation observed in customer data include:

- ▶ Use of equivalent names – TOM for THOMAS, CO for COMPANY
- ▶ Phonetic spellings – SMITH for SMYTH
- ▶ Hyphenated or compound names – NEYMAN-PEARSON versus PEARSON
- ▶ First name, last name reversals and first name, middle name reversals
- ▶ Partial matches – THE COCA COLA COMPANY versus COCA-COLA
- ▶ Typographical errors – 07151952 versus 07151925
- ▶ Incomplete or inaccurate addresses

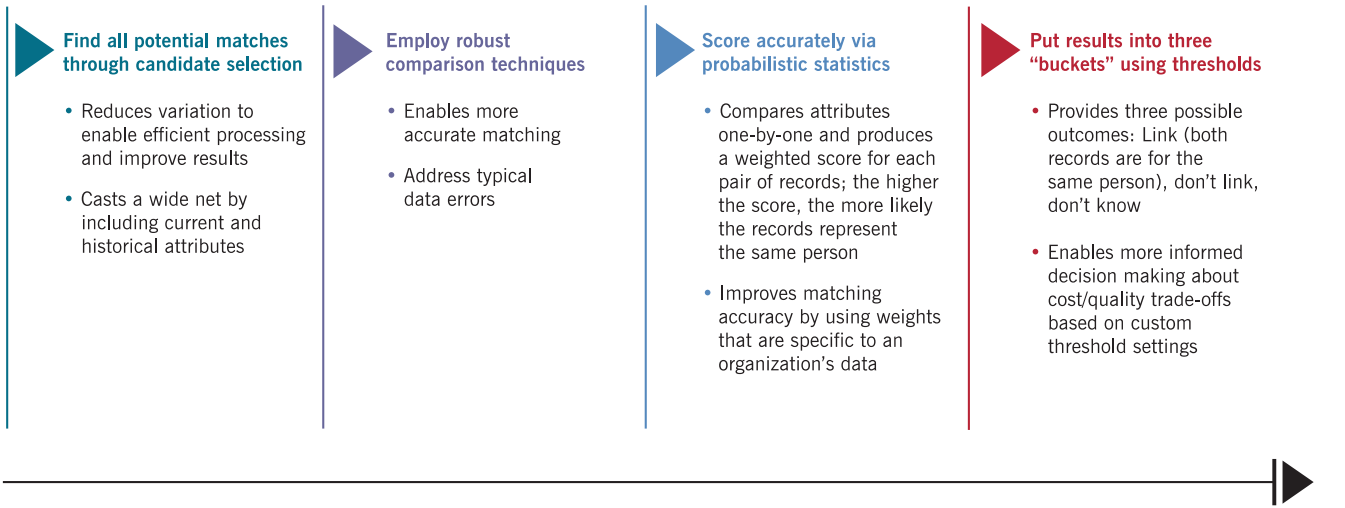
False positives arise from two primary sources:

- ▶ Matching on high-frequency values such as commonly occurring names
- ▶ Related records such as family members or inter-related businesses

The Neyman-Pearson lemma describes the optimal test statistic for this type of problem: the likelihood ratio test. This lies at the heart of Initiate's matching algorithm, which has been used to analyze and successfully match billions of customer records.

Understanding the variation in recording demographic information drives the design of candidate selection and comparison routines. Scoring and thresholding rely heavily on statistical theory. Initiate's expertise and experience in each one of these areas ensures a highly accurate and efficient matching system.

INITIATE ACHIEVES UNRIVALED MATCHING ACCURACY VIA A THOROUGH, MULTI-STEP PROCESS.



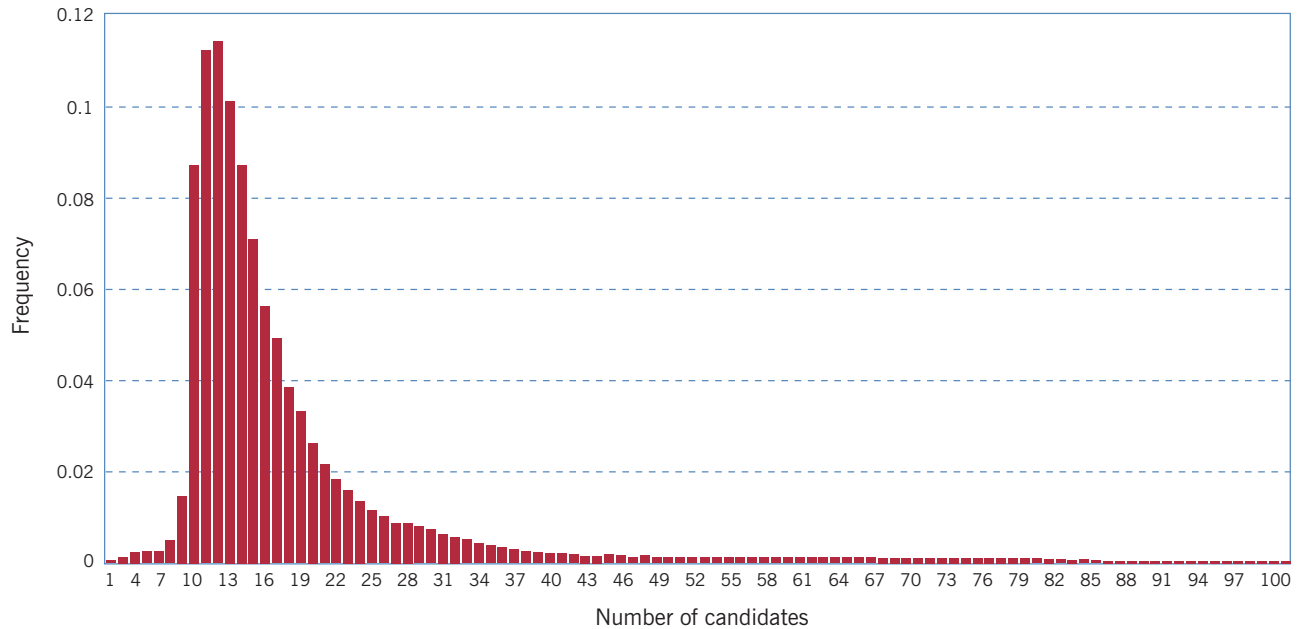
Achieving the highest levels of accuracy requires a methodical and thorough process. Initiate Systems has customized the following process to successfully match billions of records for numerous clients.

Candidate selection

The variations in recording demographic data hamper matching in two components of the matching process – candidate selection and comparison. The candidate selection process selects records from the database that are likely to match; meager candidate selection is one of the primary causes of false negatives in many systems. For each record in the database search keys are created. This enables the record to associate with other records that likely match it. Search keys are designed to be robust against the types of common errors listed above and create multiple keys for each record. These multiple, robust search keys insure that all records with a reasonable chance of matching are considered, a process known as casting a wide net. Returning to the previous example, THOMAS STEVEN JONES, search keys are created on all possible combinations, so the net would catch TOM JONES, STEVEN THOMAS, STEVE JONES, STEVEN THOMAS JONES, etc.

This approach, which is vital to achieving a low false-negative rate, creates efficiency requirements for the matching process. The following example illustrates this approach. Derived search keys were analyzed from a typical Initiate Identity Hub software installation involving approximately two million members. On average, nine separate search keys were generated for each member. Convolving the distribution of generated keys with collision statistics for those keys results in the distribution of the number of candidates considered. The following histogram illustrates the first 100 points of that distribution.

INITIATE'S WIDE NET CANDIDATE SELECTION PROCESS ENSURES THAT ALL RECORDS WITH A REASONABLE CHANCE OF MATCHING ARE CONSIDERED



For this installation, the average member would be compared to over 17 other records in finding potential matches. However, the long tail of this distribution drives the processing requirements. While the average number of comparisons per member is 17, approximately one in a thousand queries compares to 300 other records, and one in ten thousand queries requires nearly 1000 comparisons. Thus casting a wide net on even a relatively small database requires a highly efficient matching process.

Comparison functions

Comparison functions operate at the attribute level and determine the degree to which the attributes match. This step pulls together disparate data and is the foundation for building a complete record and 360° view of each customer. These can be simple binary functions, either the attributes agree exactly or not, or complicated hierarchical comparisons involving phonetic coding and edit distance functions. Based upon its extensive data experience, Initiate has developed a unique library of comparison functions to catch a wide-range of recording errors across many different types of attributes. Based upon analysis of the data, Initiate selects the right comparison function for each attribute available for matching to ensure the highest levels of matching accuracy. These functions include:

- Name comparison that:
 - Uses a comparison hierarchy for each name token, considering exact match, nickname match, name-to-initial match, and phonetic match
 - Tests all possible token arrangements between the two records
 - Eliminates anonymous values, such as TEST, before comparison

- Address and phone comparison that:
 - Standardizes the address
 - Employs intelligent parsing when the address cannot be standardized to extract useful match information
 - Tests for typographical errors
 - Recognizes and comprehends the statistical correlation between address and phone matches
- Date evaluation that:
 - Comprehends year frequency
 - Eliminates anonymous dates as determined by frequency analysis
 - Comprehends typographical errors
- SSN and other license number comparisons that comprehend typographical errors
- Email address comparison
- Credit card number comparison
- General frequency-based comparisons that can be applied to many types of attributes such as birthplace, race, gender, and key words
- General distance comparisons that can be applied to attributes such as part numbers and account numbers

A set of robust comparison functions, based upon data experience, provides a good foundation for an accurate matching system. Scoring and thresholding apply statistical theory to build from this foundation.

Scoring

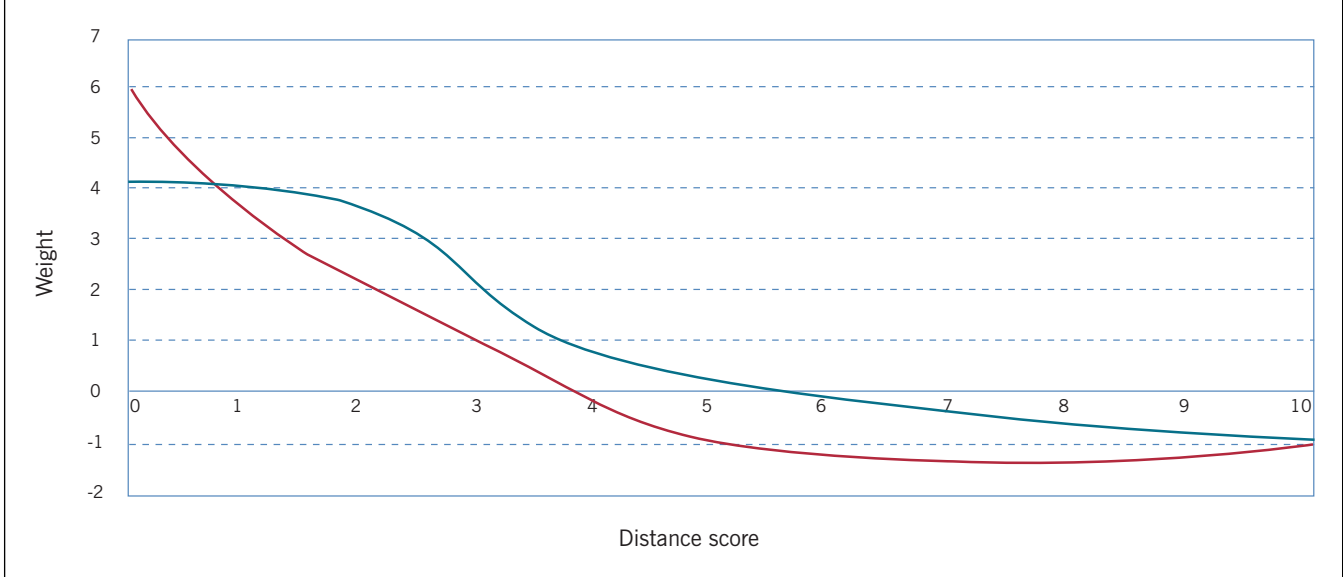
Scoring provides the mechanism for combining the individual attribute comparison results into a meaningful value. Here likelihood-ratio theory is applied. When two records are compared, the likelihood ratio is computed for the hypothesis that the records refer to the same member, to the hypothesis that they refer to different member. The theory shows how to value a match on a particular attribute to a mismatch, but it also specifies how to weight a match on one attribute to a match on another attribute. Using this approach avoids the ad-hoc nature of other matching schemes by essentially weighting each attribute's contribution to the total score by the information content of the attribute. Again, according to the theory, this test is the optimal one for this type of problem.

Any likelihood ratio test is only optimal if the probability density functions which relate our observations to our hypotheses are known. Usually, of course, they are unknown. While other approaches use heuristic estimates of these probabilities, Initiate's approach is to estimate these probability density functions based upon the particular data being processed. Thus the probability of matching on a name token "Acme" at random (i.e., under the unmatched hypothesis) is based upon the frequency of the name token "Acme" in the data file being processed. Therefore, a match on a common surname, such as "Smith," provides less valuable information than a match on a less common surname, such as "Stoker." For other comparison functions, such as edit-distance, a bootstrap technique is employed to estimate the densities. In this way, matching algorithms adapt to each particular data file – data is used to provide the best estimates possible of the underlying probabilities. This in turn provides the best test with the lowest false-positive rate.

Another advantage of this approach is that it allows general comparisons to adapt to both different attributes as well as different data. The following graph illustrates this concept applied to two different

identifiers: Tax Identification Number (TIN), and Unique Physicians Identifier Number (UPIN) from a physician-matching project. For both attributes, two values are compared using the same edit-distance function. The curves show the weight or score given for various amounts of discrepancy (zero being an exact match, one being a single character or transposition difference etc.) between the attribute values. The blue curve corresponds to TIN and the magenta curve to UPIN.

WEIGHTING DISCREPANCIES BETWEEN RECORDS ON EACH ATTRIBUTE PRODUCES A SCORE INDICATING THE LIKELIHOOD THAT A PAIR OF RECORDS REPRESENT THE SAME PERSON. THIS CHART SHOWS HOW THE SAME LEVEL OF DISCREPANCY RESULTS IN A DIFFERENT WEIGHT BY ATTRIBUTE.



For UPIN, an exact match (corresponding to a distance score of zero) yields a weight of six while an exact match on TIN produces a four. This simply reflects the fact that TINs may be shared among different physicians where UPIN is not shared (i.e. matching on UPIN provides more information that the records represent the same physician than a match on TIN). As importantly, the shape of the response curve adapts to the data as well. Thus a single character discrepancy (i.e. a distance score of one) has a large impact on the UPIN weight (dropping from six to four) but little impact on the TIN weight.

Formally, the data is used to estimate the actual likelihood ratio at each distance score. Thus the comparison function used in the Initiate process has no pre-defined shape as it transitions from full match to mismatch. The statistics of the data determine the shape.

Thresholds

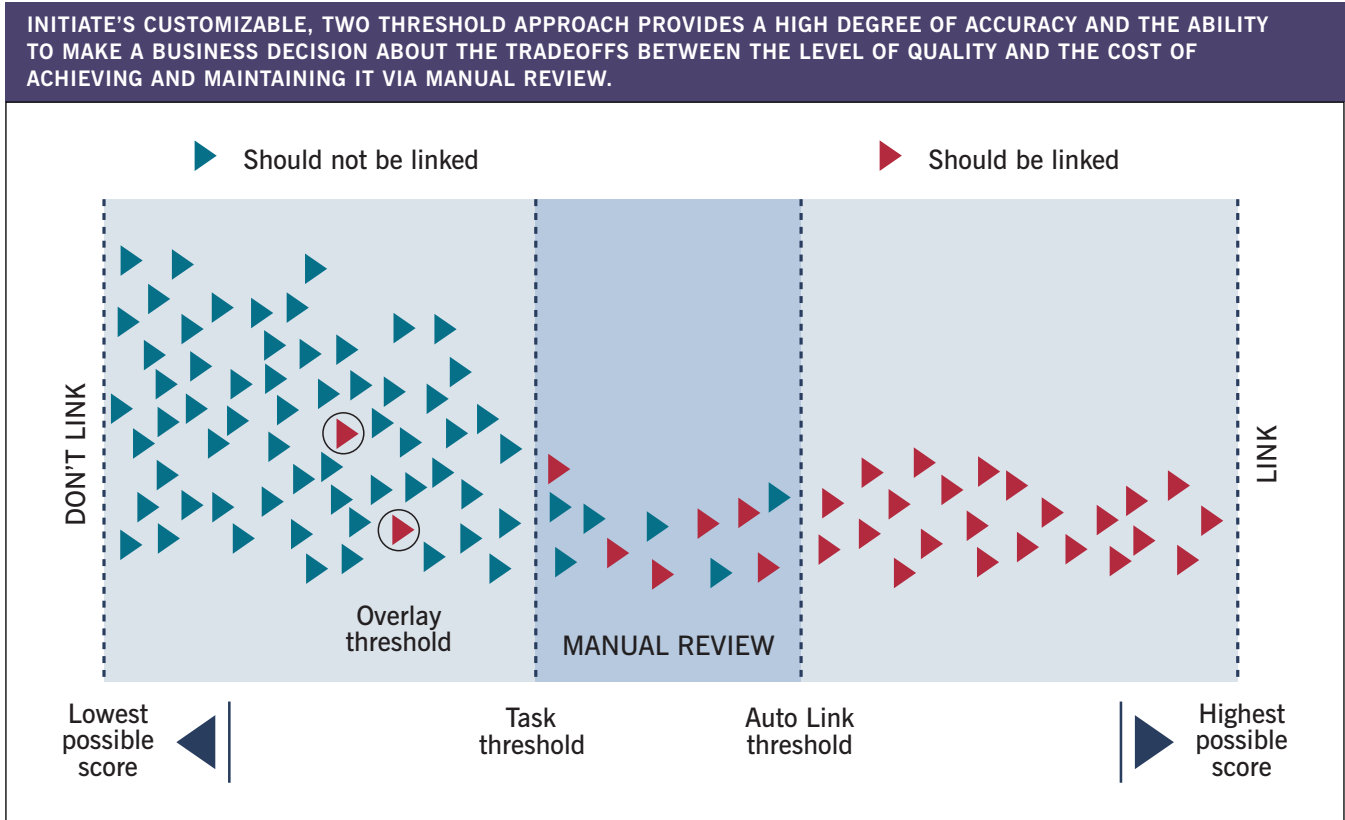
Thresholds provide the link between the statistical theory and the underlying business requirements. In the simplest example, there is a single threshold and if two records achieve a score above that threshold they are linked as referring to the same member. Conversely, if the comparison score falls below the threshold, then they are treated as referring to separate members.

In a recent installation of the Initiate Identity Hub™ software, a threshold was chosen to achieve a false-positive rate of less than 1 in 100,000 transactions (for this company, approximately ten per month). At this threshold, over 90 percent of the duplicates were linked for a false-positive rate of less than 10 percent. If it is necessary to catch more duplicates the threshold could be lowered, but the number of false positives would increase. In this example, lowering the threshold to the point where 95 percent of the duplicates are identified would mean a false-positive rate of 1 in 25,000

► WHITE PAPER

transactions. In a single threshold model, even one employing the optimal decision algorithm, decreasing one type of error necessarily increases the other.

In many business situations, this may not be acceptable. A dual threshold model, illustrated below solves this problem.



Here, two records are automatically linked if they score above the upper threshold, which is set to yield an acceptable false-positive rate. The lower threshold was chosen to yield the desired false-negative rate and records which score between the two thresholds are placed in an electronic queue for manual review (obviously, if the lower threshold, which yields the desired false-negative rate, is greater than or equal to the upper threshold, this reverts to a single threshold system).

In this instance, rather than trading performance against one type of error against performance against the other, uniform performance is traded for manual review work. The ability to set two thresholds instead of one allows the user to not compromise on either type of error but instead employ a manual review process for those records that fall into a more nebulous score range.

Only by optimizing all steps – candidate selection, comparison functions, scoring and thresholds – can a highly accurate, cost-efficient matching process be achieved. This is crucial for not only accurate data matching, but also data integrity in general and customer-centric initiatives that rely on successful data aggregation.

Evaluating data matching and linking solutions

There are many partial or related data matching solutions from other software categories (data quality, data providers/aggregators, data warehousing, for example), caveat emptor. Typically, these solutions were originally designed for some other function (data cleansing or address standardization, for example) and have been “retrofitted” for data matching. When processing a small file – typically under one million records – it is possible to “make do” with some basic data matching applications. However, as the number of records climbs above 5 million, more advanced functionality is required. When evaluating these more robust solutions it is recommended that they be evaluated against the following key attributes:

► Accuracy

- Does the provider conduct frequency-based file analysis to provide weighting and thresholds specific to a customer’s data?
- Does the provider offer complete versioning of all historical attribute values, which improves accuracy and the ability to make the correct decisions on linking data?
- Does the solution include single or dual threshold capabilities that enable tunable accuracy and automate identification and resolution of errors?
- Does the solution include proven probabilistic algorithms which score and match data across a variety of attributes using likelihood statistical theory for the highest levels of accuracy?

► Real time

- Does the solution work in real-time if required? Most solutions with a data quality legacy work in batch mode. To provide up-to-the-moment view of customer data across the enterprise, as well as prevent duplicate records from entering systems on an ongoing basis, the ability to operate in real-time is essential.

► Non invasive

- Does the solution logically link related customer data without the requirement of a common key?
- Does the solution require programming code to be added to the source systems or modification or standardization of the source data?

► Data model

- Does the solution support core data from the sources in its original form and maintain complete historical versioning?

► Task model

- Does the matching solution permit the setting of multiple matching score thresholds to help manage the quality vs. cost tradeoffs?
- Does the solution include a complete task model to prioritize and resolve data errors or ambiguous linkages?

► Security

- Does the solution offer role-based security access down to the attribute level?

► Dynamic enterprise view

- Is the solution able to deliver a configurable, 360° view of a customer, even though individual records may be sparse? Once again, an essential capability.

Conclusion

Managing identities across organizations as well as within and across business lines to support business priorities is becoming increasingly important. Existing customers in particular are increasingly frustrated by companies offering them products and services they already have – all because data integrity is so poor that many companies have no clue who many of their customers are.

The underlying cause for this confusion is the variation in customer identities, such as discrepancies in name, address, numerical identifiers and other customer unique attributes. This creates major barriers to ever achieving a consolidated view of the each customer. Ultimately, this fragmentation limits the effectiveness of customer-centric efforts such as CRM initiatives.

Efforts to manage this variation, such as "de-duping" customer files typically begin at the data warehouse level or in support of direct marketing requirements. Initially, batch processing through merge and purge tools may be acceptable, however many of these methods often overwrite existing data. Unlike data matching where records can be "unlinked", if two records are merged together and data is overwritten, it cannot be "unwritten" later.

As the need for real-time access to integrated customer data spreads across the enterprise, a flexible zero latency solution is required. Furthermore, levels of data quality offered by existing tools, though marginally acceptable for direct marketing or business analysis, is unacceptable to meet other needs that require accuracy and precision. Point of service or other requirements require comprehensive views, real-time, accurate and up-to-date data, regardless of the channel.

Most systems lack the data integrity required to support aggregation techniques since they simply were not designed with aggregation in mind. Initiate's sophisticated and accurate matching process overcomes this lack of data quality enabling organizations to finally come to grips with these mission-critical tasks.

About Initiate Systems

Founded in 1995 and headquartered in Chicago, Initiate Systems, Inc. is a leading provider of customer data integration software and services for managing customer or patient identities. Initiate helps companies that require on-demand access to accurate, up-to-the-moment customer or patient data. Initiate's solutions deliver high business value in many industries, including healthcare, financial services, hospitality and the public sector, and across many applications, including CRM and business intelligence. Initiate's award-winning Initiate Identity Hub software and its related applications deliver the highest degree of accuracy on demand, regardless of data volume. During Initiate's eight-year history, the company has analyzed billions of individual customer records on behalf of hundreds of customers. For additional information, visit www.initiatesystems.com.

Footnotes

¹ Ted Friedman, Principal Analyst, Gartner, September 2003.