ORACLE
Financial Services

# Reducing False Positives without Increasing Regulatory Risk

ORACLE WHITE PAPER | NOVEMBER 2017

ORACLE

## Table of Contents

## Introduction

False positives are the scourge of the Money Laundering Reporting Officer (MLRO)—the person responsible for protecting the reputation and security of a financial institution. Every occurrence of a client record matching a name on a sanction, risk, or PEP (politically exposed persons) register has to be investigated, and yet the review and research of false positives costs institutions time and manual effort. “Fuzzy” techniques are essential to finding inexact matches, but they often produce large numbers of records for review, and the vast majority of these will be false positives.

With some institutions swamped by the volume of false positives, the temptation to tighten match rules can be irresistible. Although this might reduce the immediate pain of so many false positives, it often increases the probability of a more insidious risk—that of false negatives.  False positives do cost time and effort, but false negatives allow criminals access to the financial system and can result in fines for both the institution and the individual MLRO—and a loss of commercial reputation as well.

This white paper examines some of the common matching techniques and advises MLROs that, rather than having to choose one from among an array of techniques that are imperfect in themselves, they can implement a broad array of technologies now incorporated in today’s most effective watchlist screening solutions.

## Achieving a Balance

Financial institutions are instructed to take a risk-based approach to AML. But the regulators have also shown that they are willing to flex their muscles if they judge that an MLRO is failing to take adequate steps to implement adequate AML procedures, including the accurate screening of clients. No screening system can produce perfect results, so the challenge facing the MLRO is to implement a solution that produces minimal false positives without increasing the risk of missing genuine matches.

With simple matching approaches, there is a direct relationship between the number of false positives and the number of false negatives: decreasing one generally leads to an increase in the other. Fortunately, there are ways of decreasing the number of false positives without increasing the risk of false negatives. The burden of false positives can be further alleviated by adopting an approach and process that focuses effort on the highest areas of risk and removes wasted effort. This paper addresses some of the strategies available to MLROs, including:

- Match Keys – the key to matching?
- "Fuzzy" techniques – when and how to use them.
- Probabilistic and deterministic matching to increase accuracy and precision.
- Data-driven match rules – tuning rules to make the best use of your data.
- Decision support – how to review matching once, not again and again.

***False Positive:*** *A client record that screening software incorrectly identifies as a match to a record on a sanction, risk, or PEP register.*

***False Negative:*** *A client record that screening software fails to identify as a potential risk, despite the appearance of a relevant record on a sanction, risk, or PEP register.*

### Match Keys

A match key is an alpha-numeric combination that is generated from a complete record, often comprising parts of some of the key elements, such as postcode and name. Match keys have long been used to de-duplicate mailing lists for marketing purposes. They offer the convenience of fast, high-volume matching but lack the sophistication needed for AML operations and generate large volumes of false positives and false negatives.

Although they are not usable on their own for AML purposes, sophisticated match keys can be used to group records for more detailed comparison with fuzzy techniques.

### Fuzzy Is What Fuzzy Does

"Fuzzy matching" is a phrase used for any technique that allows for the identification of non-exact matches. For example, matching the first name Elizabeth with Elisabeth requires some form of fuzzy matching. In this case, any of the common fuzzy techniques would have found a match, but each of them has strengths and weaknesses. Some of the more common fuzzy matching techniques are evaluated here.

#### Soundex

Soundex is an algorithm for encoding a word so that similar-sounding words produce the same encoded answer. A Soundex code consists of a letter followed by three digits. The first letter of the word is retained with subsequent consonants being converted to numbers according to a scheme which groups letters that are most commonly confused (for example, the letters B, F, P and V are all encoded as 1). Vowels, some consonants (the letters H, W, and Y), and repeated letters are ignored, and zeroes are added if necessary to complete the full Soundex code.

We can see from the examples in the Table 1 below that the names Elizabeth and Elisabeth would return as a match when using Soundex, but it's also clear that Elizabeth would not be matched to other variants, such as Elsa or Betty, when using this method. Soundex can also prevent matching some records that are good matches to the human eye—take Christian and Kristian, for example, which do not match using Soundex.

Soundex works on the assumption that the first letter of a word is most likely correct, so it fails to identify matches when this is not true. It also lumps together too many false matches, such as Christopher and Christine, to be entirely reliable. Additionally, it should be remembered that the Soundex method of encoding was developed mainly for words in the English language and performs less well for words with other origins. The names Schultz (S432) and Schulz (S420), for instance, do not match when using Soundex.

**TABLE 1. SOUNDEX SAMPLE MATCHES**

| First Names | Soundex Code |
|---|---|
| Elizabeth and Elisabeth | E421 |
| Elisa and Elsa | E420 |
| Beth and Betty | B300 |
| Christopher and Christian | C623 |
| Chris | C620 |
| Christine and Christina | C623 |
| Kristian | K263 |
| Kristin | K623 |

### Metaphone

Metaphone is a phonetic algorithm based on the sound that words make in the English language. It was originally developed to overcome the shortcomings of Soundex and works by coding consonants (or combinations of letters), depending upon their position in the word. For example, the letter C can be coded as X, S, or K, based on where it appears and what letters surround it. The combination TH is encoded as 0. Unlike Soundex, Metaphone results can be of any length.

As with Soundex, the names Elizabeth and Elisabeth would return a match using Metaphone, but Elizabeth would not achieve a direct match to other variants, such as Elsa or Betty. However, Elisa and Elsa do both produce a Metaphone value (ALS) that matches the start of the Metaphone for Elizabeth's longer string (ALSP0).

Where Christian and Kristian were not matched by Soundex, we can see that these successfully match using Metaphone. The algorithm has also correctly identified Christopher and Christine as significantly different and not matched them. However, Metaphone also suffers (perhaps more so) from being built for the English language. Although there are variations of the Metaphone algorithm for different languages, the standard routine can struggle with names from other source—again, the names Schultz (XLTS) and Schulz (XLS) do not match.

**TABLE 2. METAPHONE SAMPLE MATCHES**

| First Names | Metaphone Code |
|---|---|
| Elizabeth and Elisabeth | ALSPO |
| Elisa and Elsa | ALS |
| Beth | PO |
| Betty | PT |
| Christopher | KRSTFR |
| Chris | KRS |
| Christian and Kristian | KRSXN |
| Christine, Christina, and Kristin | KRSTN |

### Edit Distance

Another way of comparing two names (or other text data) is to look at the number of character differences between two fields, often described as a "spelling comparison" or "edit distance comparison." Different versions of the routine may have distinct scoring algorithms, but the principal is always the same—determine how many edits (character changes) would need to be made to one record in order to make it identical to the other record.

Identical names produce an edit distance score of 0, and those that are different (but close) deliver a low score. In this scheme, the pair Elizabeth and Elisabeth achieves a score of 1 and would generally be considered a positive match. The routine is not good, however, at identifying the correspondence of name variants such as Elizabeth and Beth.

Another general weakness of the edit distance comparison is that a raw score does not take into account the length of the text being compared. A comparison of J to Jo would score 1—a closer match than Christopher and Kristopher, which score a score of 2—but most people would consider the second example a better match. Some variations on the edit distance comparison deal with this weakness by allowing users to weight matches according to the length of the data involved.

**TABLE 3. EDIT DISTANCE SAMPLE SCORES**

| First Names Pairs | Edit Distance Score |
|---|---|
| Elizabeth versus Elisabeth | 1 |
| Elisa versus Elsa | 1 |
| Christian versus Kristian | 2 |
| Elizabeth versus Beth | 5 |
| Kristian versus Kristin | 3 |
| Schwartz versus Schwarz | 1 |

### Equivalence

The Equivalence algorithm uses reference data to identify pairs of records that are logically equivalent. This gives the Equivalence routine the ability to match Elizabeth to Elisa, Elsa, Beth, and Betty, as well as to Elisabeth. It can deal equally well with names in any language and can also recognize variations of names rooted in any country.

Provided with the appropriate reference data, the Equivalence algorithm can also deal with legal entity name matching, but it is entirely dependent upon this data and cannot deal with simple typographical errors unless they are listed.

### Other Text Comparisons

The match algorithms discussed thus far in this paper have all focused on comparing single words, but other routines are equally capable of processing data items containing multiple words (full names and legal entity names, for instance), such as the following:

- **Word edit distance.** This method operates in a similar way to the character edit distance comparison (discussed above), but it counts the words that disagree between the two fields. Comparing *Joseph Andrew Cole* to *Joseph Cole* with the character edit distance method would score a 6 (a poor match), but the word edit distance for these strings is just 1.
- **Longest common substring and longest common phrase.** The "longest common substring" method produces a score based on a character count or on the length of the longest contiguous sequence of characters in the records being compared (represented as a percentage of the entire phrase). For example, the pair *Inventum Securities* and *Inventum Securities Ltd* scores 19 or 83 percent). The "longest common phrase" method produces a score based on the longest string of words that concurs (*Inventum Securities* and *Inventum Securities Ltd* scores 2 or 67 percent).
- **Word match.** This method returns a count of words (or percentage of the number of words) that occur in both records, regardless of their position. For example, *David Sheldon Turner* and *Turner David Sheldon* would score 3 or 100 percent.

## Increasing Accuracy and Precision

Most customer matching technology was originally developed for marketing purposes, where the key driver has been to maximize matches, and false positive matches have been an accepted side effect. The screening of client records for regulatory purposes demands a much higher level of accuracy to reduce false positives without increasing the risk of false negatives.

### Probabilistic and Deterministic Matching

Some vendors of matching software describe their software as either probabilistic or deterministic and make claims about the virtues of one over the other. Probabilistic matching compares two records and returns a percentage indicating the likelihood (according to the algorithm) of a match. Probabilistic matching systems weight scores based on the frequency and uniqueness of data and require (or allow) little tuning and configuration.

Deterministic matching uses a combination of comparisons and business rules to determine when records match—a rule might, for instance, require a match on the client name and year of birth. The result from the comparison of records is either a match or a no-match according to business rules that have been defined by the user.

Both probabilistic and deterministic matching have their virtues and the two approaches can be used together to offer an optimal solution. Probabilistic comparisons, using fuzzy techniques, can be used as part of a deterministic scheme that gives the user full control. Identified matches can be ranked according to the user's business rules and potential matches can be reported as well as definite ones.

### Adequate Data Preparation

One of the most important (but often overlooked) methods of reducing false positives and regulatory risk is to ensure accurate preparation of data sources ahead of screening. Instances such as overfilling of name data, misspellings, multiple date-of-birth formats, non-standard country information, and nonstandard name constructs can often persist in both customer and reference data sources.

This situation is compounded further when customer data is held in multiple writing systems, requiring a process for converting non-Latin customer data into the Latin form as part of data preparation. Techniques such as transliteration can be used to convert from one writing system to another using character-level rules, but more complex languages (such as Arabic) will require the use of transcription and variant matching to accurately identify all potential name equivalencies.

Techniques such as profiling, auditing, transformation, and text analysis can be used to validate data, remove white spaces and possibly erroneous characters, or split a single name field containing multiple attributes into a number of fields. These capabilities enable organizational data to be optimized to match rules.

## The Matching Process

The recommended approach for MLROs is a method that enables business users to define the rules that determine which records are considered to be matching. A granular comparison of records, along with the use of match patterns, takes the best of multiple methods discussed above but still incorporates user-defined rules. The following four steps explain the method in detail.

1. On their own, Match Keys do not offer the precision and accuracy required for client screening; however, they can be used to provide high-level matching. This enables similar records to be grouped for further comparison. It is good practice to use multiple Match Keys (built from different components) to ensure matching records will be clustered together.
2. All records that share a common Match Key can then be clustered together. When multiple keys have been generated for each record, the record can appear in more than one group. This significantly reduces the chances of false negatives that can occur when records don't meet in a cluster.
3. Each record in a cluster can then be compared, in detail, to every other record in the cluster. Comparisons are carried out at the field level, with each comparison producing a result according to a defined business rule or a probabilistic score. Together, the results for each comparison build a match pattern for the whole record comparison.
4. The resulting match pattern is then validated against the user-defined business rules. These detail what combination of field level results can be considered a definite or potential match. The results can also be ranked according to the institution's risk profile, and potential matches can be flagged for manual review according to their ranking.

By tuning these rules to their own data and risk profile, financial institutions are able to control the screening process and achieve a high degree of precision and accuracy with minimal false positives.

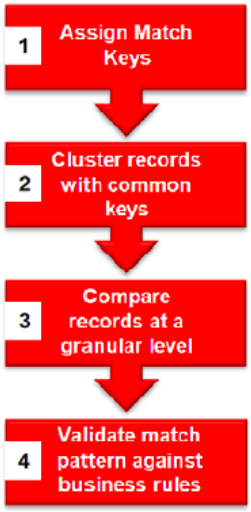


Figure 1. A four-step method combining the benefits of using multiple match keys, clustering, granular matching and user-defined business rules to minimize false positives.

## Data and Risk-Driven Match Rules

The high-level money laundering threat is the same for all financial institutions and the reference lists available (from both government sponsored bodies and commercial suppliers) are common, yet MLROs at each institution face unique challenges—how to manage their own data and risk profiles. A “black box” solution, where the rules are completely predefined, does not allow for these factors to be taken into account. Instead, MLROs should look for a solution that includes a set of proven, preconfigured rules but is also easily tuned to meet the particular data and needs of their institutions.

Data profiling and analysis delivers rapid insight into the population levels and quality of every available attribute. This discovery exercise can help to identify the fields that can be used to screen client records accurately against the prescribed risk register. It will also reveal any requirements for preparation of the data prior to or during matching.

The matching process itself should tolerate some level of data transformation. For instance, when comparing business names, the solution should be capable of ignoring less significant words, such as business type terms including “& sons, Ltd., Limited, and plc.” The solution should also allow for words to be standardized for matching. More complex data preparation (such as parsing names to split them into title and first and last names) requires a data preparation phase before the match process.

Consequently, a complete client screening solution should have the ability to profile, analyze, and transform data and should feature a sophisticated matching engine as well.

## Decisions, Decisions

Every client record that is matched to a record on the risk register needs to be reviewed and a decision needs to be made. Is your client the same person recorded on the register? The majority of matches will be false positives so, once they’ve been reviewed, there should be no need to look at them again unless something changes. The ability to flag a client record as “decided” can make an enormous difference in the amount of time and effort that is required for manual intervention, typically reducing costs by as much as 70 percent. However, few screening solutions provide support for the review of potential matches and even fewer remember the decisions that are made.

When a case management tool is integrated with the screening application, the result is improved efficiency. This tool enables the MLRO to consolidate multiple entries of the same individual or entity and review all potential matches between a single customer record and one or more watch-list entries. Whether this is included as part of the screening solution or not, the MLRO must be able to record an audit trail about every manual decision. The audit trail must include a note of who made the decision, when it was made, the information they had available from the client data and risk register, plus any supporting notes for the decision. The case management tool should incorporate an automated audit trail and the ability to hand over or escalate cases (with case notes suppressed where appropriate to protect customer confidentiality). It is also critical that the tool supports automatic escalations, which would occur against predetermined intervals such as high risk alerts—that could potentially expose institutions to financial loss or reputation damage.

## Exposing Multiple Identities

The use of multiple identities is common in the criminal world and Al-Qaeda's own training manual requires its operatives to use false identities to hide their terrorist activities. Exploiting variations of a criminal's real name is perhaps the simplest way of acquiring a new identity. Typical approaches are to use name variations (for example, Robert might use the names Bob, Bert, or Bobby), or to switch the order of names (for example, Thomas Howard becomes Howard Thomas and James Richard Smith becomes Richard James). Other data, such as dates of birth, may also be simply manipulated by transposing digits, so that 12/11/1956 becomes 11/12/1965.

Commercially available risk registers such as C6, Dow Jones Solutions for Risk and Compliance, World-Check, WorldCompliance, and others often include multiple variations of names and known pseudonyms. However, MLROs need a reliable way of applying these to their own customer data, as well as the ability to identify manipulated identities that have not previously been recorded.

The best approach is to use sophisticated fuzzy matching algorithms, which expose identity-masking techniques such as transpositions, spelling variations, alternative name variations, and provide transliterations of foreign script names (such as names in Arabic, Chinese, or Cyrillic).

### Minimizing False Positives: 6 Points to Remember

- False positives cost time and money; false negatives could cost your reputation.
- It's possible to reduce the number of false positives by deploying appropriate matching technology and business rules.
- The cost of reviewing false positives can be reduced by as much as 70% by using a review tool that records an audit trail and learns from the decisions that you make.
- Match keys do not provide the matching accuracy that is required for AML screening, but they should be used to cluster records for detailed comparison.
- Fuzzy matching techniques can help to identify matches, but none of them should be used exclusively. Combining them into a match pattern allows for greater precision and accuracy.
- A rules template provides a fast start position, but it is essential that to be able to tune the rules for your data and risk profile.

## Conclusion

In this white paper, we have discussed the different techniques and highlighted the strengths and weaknesses of each method of matching in isolation. Used on their own, each technique will catch only those individuals or entities that fit the criteria being applied. In an effort to broaden the search method to protect against all threats, the temptation can be to loosen the search criteria. This inevitably results in a greater number of false positives.

Oracle's Customer Screening solution offers more than two hundred proprietary fuzzy logic screening algorithms, including all of the techniques mentioned in this paper and many others. The advanced data-matching algorithms have been tested in open trials against other technologies and proven to be the most accurate and effective customer screening algorithms in the industry.

**Oracle Corporation, World Headquarters**
500 Oracle Parkway
Redwood Shores, CA 94065, USA

**Worldwide Inquiries**
Phone: +1.650.506.7000
Fax: +1.650.506.7200

ORACLE

CONNECT WITH US

blogs.oracle.com/financialservices

facebook.com/oraclefs

twitter.com/oraclefs

oracle.com/financialservices

Integrated Cloud Applications & Platform Services



White Paper
Reducing False Positives without Increasing Regulatory Risk
November 2017