Whats in a name?

The name matching you need, when the implications of not identifying a match are too great

The Theory and the Science of Name Matching Technologies

ABOUT SEARCH365 AND OVERWATCH

Search365 provides software solutions for text analytics, information retrieval, digital forensics, and identity resolution in over forty languages. Search365 also provides our clients with professional and managed services in the areas of business intelligence, advanced analytics, e-discovery, social media monitoring, financial compliance, and other enterprise applications.

Our Over Watch and IdentifAI platform combines enterprise search with the most advanced name matching and identity resolution capabilities previously only available in limited national security use cases. Our linguistics team is at the forefront of applied natural language processing using a combination of statistical modeling, expert rules, and corpus-derived data.

www.search365.ai.

1. THE CHALLENGE

The need for name matching has moved far beyond de-duplicating mailing lists to save a few dollars on mailing one less catalog. Today, in financial compliance, law enforcement, and national security, the costs are much higher. The costs of a false positive—a wrong match— are the time and money wasted by people reviewing the erroneous matches, inconveniencing innocent people, or embarrassing customers. The costs of a false negative—a missed match— are the increased risks of reputation damage, regulatory fines, and known risky individuals going unchecked.

In the most basic of needs, the ability to tie data together across multiple data sources to obtain a 'single customer view' is an old and challenging problem for business's, small and large, to solve. Beyond this and the financial cost, there are many scenarios where a missed match could even put lives at risk.

We will explore four name matching methods which are applied to these high-stake situations: common key (e.g., Soundex), lists of name variations, edit distance, and statistical similarity. We will examine the strengths and weaknesses of each method with respect to matching names written in both English and other languages, and conclude with the current "best practice" among name matching technologies.

2. PROBLEM DEFINITION: NAME VARIATIONS - HOW AND WHY NAMES DIFFER

Name matching challenges are more complex than almost any other data type. The variability of names coupled with language norms that vary globally in many languages, present a major challenge. It's no longer just matching Jimmy L. Smith to James Larry Smith, but "عبد الراشد" with "Abdul Rasheed" or "Zhang Jing-Quan" with "Jingquan Zhang."

The vast majority of name matching variations that occur within a language and across languages can be categorized as follows.

- **Typographical errors** - A slip of the finger at the keyboard causes transposition of characters, missed characters or other similar errors. (e.g., "Htomas" or "Elizbeth")
- **Phonetic spelling variations** - Some names simply sound alike, but are spelled differently. Neglecting to confirm spelling produces errors. (e.g., "Cairns" vs. "Kearns" vs. "Kerns"; or "Smith" vs. "Smyth")
- **Transliteration spelling differences** — Multiple transliteration standards or "approximate" transliterations from a non-Latin script to English lead to multiple spelling variations. In the case of Arabic to English, Arabic has many consonant sounds which might be written with the same English letter, or Arabic vowels may be expressed more than one way in English, giving rise to many spelling variations. (e.g., "Abdul Rasheed" vs. "Abd-al-Rasheed" vs. "Abd Ar-Rashid")
- **Initials** - Sometimes all name components are spelled out, other times initials are used. (e.g., "Mary A. Hall" vs. " Mary Alice Hall" vs. "M.A. Hall")
- **Nicknames** - In some cultures, nicknames are numerous and may be often used in place of a person's formal name (e.g., "Elizabeth", "Beth", "Liz", and "Lisbeth")
- **Re-ordered name components** - The order of family name and given name may appear swapped due to database format or ignorance of cultural naming convention. (e.g., "John Henry" vs. "Henry, John"; or "Tanaka Kentaro" vs. "Kentaro Tanaka")
- **Missing name components** - Sometimes a middle name or patronymic (personal name derived from ancestor's name—e.g., Olafsson = "son of Olaf") may be absent. (e.g., "Abdullah Al-Ashqar" vs. "Abdullah Bin Hassan Al-Ashqar"; or "Philip Charles Carr" vs. "Phillip Carr")

- Missing spaces - Some names are commonly written with spaces in different places, both in common English names (e.g., "Mary Ellen", "Maryellen", and "Mary-Ellen") and those less common in English (e.g., "Zhang Jing Quan" and "Zhang Jingquan").
- Truncated name components - (e.g., "Mcdonal", "Stev")
- Names in different languages - Names from languages using different writing systems can be notoriously difficult to match against English representations of the names. Here is just one name spelled in English, Russian, simplified Chinese, and traditional Chinese, respectively: "Mao Zedong", "Мао Цзедун, 毛泽东 or 毛澤東").

Let's consider further the reasons for names to vary, particularly focusing on the case when a name is written in a language other than its native one.

3. IN-DEPTH: FOREIGN NAMES IN ENGLISH

Transliteration is conversion of a name from its native script to another one, based on the spelling of the original name. The Japanese girl's name あずさ is transliterated to Latin script substituting "a" for あ,"zu" for ず, and "sa" for さ to produce "Azusa" as its English transliteration.

Transliterated names present a number of unique challenges to a name matching system. Transliterated spelling variations, re-ordered or missing name components, and mis-segmentation of names occur frequently with names that are natively written in a non-Latin writing system.

3.1. Transliterated spelling variations

Arabic is a good example of a language that contains many more sounds than English, which often produces ambiguity in mapping names from Arabic to English. Arabic emphatic consonants, for example, are difficult to express with just English characters when both Arabic characters have an "s" sound.

Regional variation also influences Arabic name spelling in English. A particular Arabic letter may be pronounced with a "k" sound in Libya, but a "g" sound in Egypt. Should one then write "Kaddafi" or "Gaddafi"?

Myriad Arabic transliteration standards exist, and more variations occur when an English speaker or a French speaker make a "best guess" at the spelling of a name heard. Should the "sh" sound be spelled "sh" (as in English) or "ch" (as in French)?

Other characteristics of Arabic, such as the assimilation of a definite article (a!) with a following consonant also produce transliteration variations. In the name عبد الرشيد do you show the phonetic result of the assimilation of the "r" between "al" and "Rashid" as in "Ar-Rashid" or do you transliterate true to the printed characters "al Rashid" regardless of the pronunciation?

3.2. Reordered name components

In East Asia (as well as in Hungary and Sri Lanka), the surname precedes the given name. But when translated into English, the order of the names is sometimes (but not always!) reversed to reflect Western conventions. For example, the Korean actor "Park Sang-Myun" (surname "Park") might also see his name written as "Sang-Myun Park" in a Western context.

3.3. Mis-segmentation of name components

Chinese and Korean are a particularly tricky when it comes to segmentation of name components. Given names in Chinese are frequently written with two characters. Should the transliteration put those two characters together (Zhang Jingquan)? Hyphenate them (Zhang Jing-Quan)? Or, write them as two words (Zhang Jing Quan)? Furthermore, where do the family and given names begin and end when each may be one to three characters long?

4. IN-DEPTH: FOREIGN NAMES IN OTHER LANGUAGES

4.1. Foreign Names from a Chinese Perspective

Foreign names are written in Chinese using characters which approximate the sounds of the foreign name or the meaning of the foreign name.

Microsoft in Chinese is written as 微软 (wei ruan), which character-by-character means “small, micro,” and “soft”. On the other hand, Coca Cola is 可口可乐 (Ke Kou ke le) a transliteration reflecting the English pronunciation.

When transliterating Western names, an individual must approximate the sounds in the original language, which means that there can be significant variation between Chinese-speaking regions. For example, the actress Natalie Portman's name is spelled variously:

你大里。报文 (Ni da li. Bao wen) in Hong kong

努塔里。博特曼 (nuo ta li . bot e man) in china

努塔里。博曼 (nuo ta li . bo man) in Taiwan

The challenges in Chinese, and other languages using ideographic characters, are compounded due to the fact that many individual characters have homophones. These six characters 陆 禄 路 鹿 逯 卢 are all pronounced "lu", but have completely different meanings. When transliterating names, any identically pronounced character may be substituted!

4.2. Foreign Names from a Japanese Perspective

In Japanese, foreign names are written phonetically using the katakana script, which has fewer consonants than English, fewer vowel sounds (just "ah", "ee" "oo" "eh" "oh"), and no diphthongs. Consequently, foreign sounds which don't exist in Japanese, can map to more than one combination of katakana, resulting in spelling variations, particularly for lesser known names.

A more unusual name variation complication involves Korean, Japanese, and Chinese names written with Chinese ideographs. Both Korea and Japan borrowed Chinese ideographs from China in ancient times and at different times, starting as early as 100 BC in Korea or the 5th century in Japan. Thus, the Chinese characters used in Korean and Japanese differ from the characters used in modern Chinese. These differences may be slight or quite dramatic. Thus a Korean or Chinese name written with Chinese ideographs may appear in Japanese using the Japanese variant of the original character used in Korean or Chinese.

Take the name of the Chinese film director Zhang Yimou. His name in simplified Chinese (张艺谋) or traditional Chinese (張藝謀) looks yet different when written with the Japanese version of the same Chinese characters, kanji (張藝謀).

And when pronounced in Japanese, Zhang Yimou sounds like Chou Imou, which is how it's pronounced when written in katakana チャン・イーモウ).

5. SURVEY OF NAME MATCHING METHODS

Focusing just on how to match names when the spelling varies, there are four basic approaches - plus a hybrid mix of the four - used by most name matching solutions:

- **Common key method** - These methods reduce names to a key or code based on their English pronunciation, such that similar sounding names share the same key. A well-known common key method is Soundex. For example, these names share the code C530: Cyndi, Canada, Candy, Canty, Chant, Condie.

- **List method** - This method attempts to list all possible spelling variations of each name component and then looks for matching names from these lists of name variations. For example: One system produced 3,024 possible transliterations of this Arabic name "j " since each separate name component alone has several variations. Here are the first five and last five variations.

1. abdal-rashid
2. abdal-rashide
3. abdal-rasheed
4. abdal-rashiyd
5. abdal-rachid

...

3020. 'abd-errshiyd

3021. 'abd-errchid

3022. 'abd-errchide

3023. 'abd-errcheed

3024. 'abd-errchiyd

- **Edit distance method -** This approach looks at how many character changes it takes to get from one name to another. "Cindy" and "Cyndi" have an edit distance of 1 since the "i" and "y" are merely transposed, whereas "Catherine" and "Katharine" have an edit distance of 2 as the "C" turns into a "K" and the first "e" becomes an "a."
- **Statistical similarity method** - A statistical approach takes hundreds, if not thousands, of matching name pairs and trains a model to recognize what two "similar names" look like so that the model can take two names and assign a similarity score.
- **Hybrid methods** - Hybrids use a combination of the above methods.

5.1. Pros and Cons of the Common Key Method

Many methods take a similar approach to Soundex, including Metaphone and Double Metaphone. These methods use phonetic algorithms which turn similar sounding names into the same key, thus identifying similar names. Metaphone expands on Soundex with a wider set of English pronunciation rules and allowing for varying lengths of keys, whereas Soundex uses a fixed-length key.

Double Metaphone further refines the matching by returning both a "primary" and "secondary" code for each name, allowing for greater ambiguity. In addition, instead of being tied to English pronunciation of characters, it attempts to encompass pronunciations of other origins such as Slavic, Germanic, Celtic, Greek, French, Italian, Spanish, and Chinese.

For example, Double Metaphone encodes "Smith" with a primary code of SMO and a secondary code of XMT, while it tags "Schmidt" with a primary code of XMT and a secondary code of SMT. That the names share a primary and secondary code of XMT indicates a degree of similarity between the names which Soundex perhaps overstates and which Metaphone misses.

Name	Soundex Key
Smith	S53O
Schmidt	S53O

Name	Metaphone Key
Smith	SMO
Schmidt	SXMTT

While the common key method is fast to execute and has good recall, the precision suffers. Manual inspection of a few names reveals the precision issues.

These names share the Soundex key H245: Haugland, Hagelin, Haslam, Heislen, Heslin, Hicklin, Highland, Hoagland

Metaphone does a better job than Soundex, encoding the above names with different codes except for the very similar pairs Haugland/Hoagland and Heislen/Heslin.

Name	Metaphone Key
Haugland	HKLNT
Hagelin	HJLN
Haslam	HSLM
Heislen	HSLN
Heslin	HSLN
Hicklin	HKLN
Highland	HFLNT
Hoagland	HKLNT

For cases where name similarity is being scored against pairs of names in different scripts—for example Korean hangul vs. English—the name must first be converted to Latin characters, which potentially introduces more errors to the comparison.

Particularly in languages such as Japanese where pronunciation is ambiguous, converting first to the Latin script can introduce fatal mistakes. Take these common Japanese female names:

澄子 can be correctly pronounced Junko or Sumiko.

幸子 can be correctly pronounced Yukiko or Sachiko.

5.2. The Weakness of the Common Key Method in Matching Across Scripts

As mentioned earlier regarding foreign names in Japanese, transliteration—to or from a Latin-based language—produces many possible variations since sounds in one language have to be approximated. Thus, adding the variations introduced from transliteration to the already difficult task of finding similar names further lowers precision.

عبد الرشيد is being evaluated against abdal-rachid, but the transliteration of عبد الرشيد produces Ar-Rashid, will the names come back as a match—as they should? The answer is No.

Name	Soundex Key	Metaphone Key
abdal-rachid	A134	ABTLRXT
Ar-Rashid	A623	ARRXT

5.3. Pros and Cons of the List Method

The list method of trying to generate every possible name variation has a couple of obvious drawbacks. Name variations which are not in the list will not be found as matches, and perhaps an even greater issue is that of speed and size. Since multi-part names–particularly non-English names–generate an exponential list of variations, searching through these lists takes time.

Given a name with just three components and 20 possible variations per name, the number of possibilities can be 203 (=8,000), a very large search space. There are further challenges with the list method – how do you score matches when one of your 8,000 query variants matches more than one name in the database? It is also difficult to handle other types of variation, like nicknames, initials, and titles, without expanding the search space even more.

A benefit of the list method is that it is simple to maintain. When a user complains about a missed match, it's easily added to the name database. However, easy maintenance may not be enough to offset the decreased speed, which is critical in high-throughput industries, such as banking and finance which are governed by Know Your Customer and Anti-Money Laundering regulations.

5.4. Pros and Cons of the Edit Distance Method

Methods which look at the character-by-character distance between two names include the Levenshtein distance, the Jaro–Winkler distance, and the Jaccard similarity coefficient. These approaches look at some combination of two factors (1) the number of similar characters and (2) the number of edit operations it takes to turn one name into the other—the operations being, insert, delete, and transpose.

Although these comparisons are quick, they do not capture linguistic nuance. All edits are given the same weight. Thus changing "c" to "p" is weighted equally as "c" to "k" although in English the latter substitution might more clearly indicate a similar name, as in "Catherine" vs. "Katherine." Further, a one-to-many character mapping is not possible, as in the case of the Arabic character "sheen" (j) which is frequently mapped to "sh" in English.

And, just as with the common key method, a non-Latin script name must first be transliterated to Latin script before the comparison can be executed, as explained in the discussion of

"The Weakness of the Common Key Method in Matching Across Scripts".

5.5. Pros and Cons of the Statistical Similarity Method

A statistical model that has been trained on thousands of pairs of matching names offers high accuracy and the ability to directly match names written in different languages without first transliterating names to Latin script. This method has a higher barrier to entry, as collecting the matching names requires significant resources, but the accuracy may be well worth the effort.

A downside is the slowness of execution. A system only using the statistical method to sift through millions of names to look for matches may be too slow to be feasible in high-transaction environments.

5.6. Pros and Cons of the Hybrid Method

Hybrid approaches acknowledge the strengths and weaknesses of the preceding four approaches and attempt to backfill weakness in one approach with the strength of a different approach.

A hybrid approach combines two or more of the above approaches, such as the common key method with the statistical method. In a first pass—taking advantage of the common key method's speed and high recall—the candidate pool is quickly winnowed down to a smaller, likely set of matches. Then a second pass over the culled down list uses a high-precision statistical method to filter the highest scoring matches to the top, making fine-grained distinctions between different matches.

Compared to the common key method alone, accuracy is greatly improved by this hybrid method. Instead of being locked into a coarse comparison of derived keys (for better or worse), the second pass of the hybrid approach takes a fresh look at the original names in their original scripts before scoring their similarity.

This hybrid method also avoids the weaknesses of the list approach by not relying on mass generation of name variations, but instead, uses (via the statistical model) the linguistic variations of names in each language. This linguistic knowledge of name variations also gives the hybrid approach an edge over the edit distance method, which cannot directly compare names in different scripts.

The result is a fast, accurate, name matching algorithm.

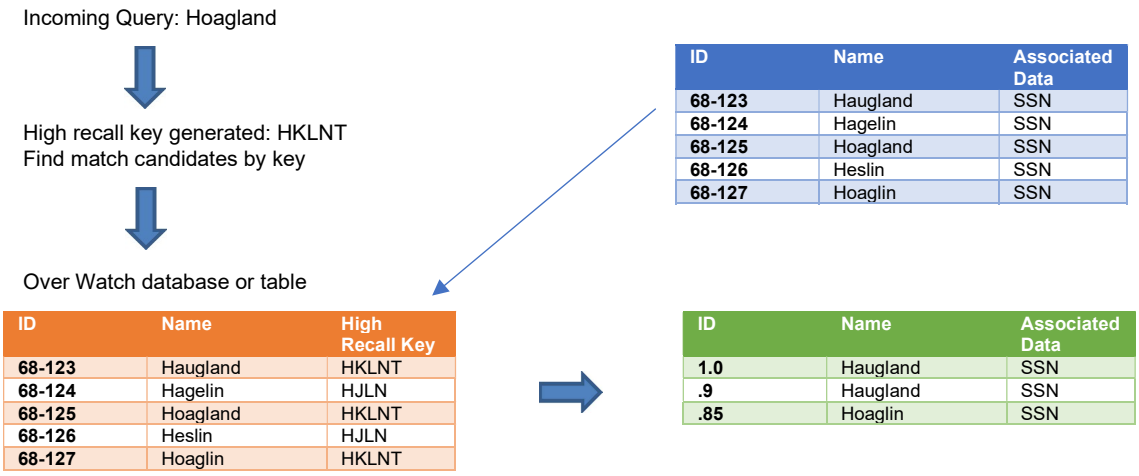
6. A HYBRID SOLUTION: OVERWATCH

Over Watch is one example of the name matching hybrid algorithm described above. Over Watch provides an API so that it can be easily integrated into any name matching application whether for financial compliance, watch list monitoring, law enforcement, or other purposes.

The name indexer goes a step further than merely combining algorithms. To handle name phenomena such as missing name components, it compares all available components looking for the best alignment of components, and then scores the degree of match, to give the end-user an appropriate degree of confidence in the match.

6.1. How It Works

With a new database, or as records are added or changed, OverWatch runs a set of common key algorithms and stores the "High Recall Keys" of each name in the database. Then when an actual query comes in, the queried name's High Recall Keys are generated and database names sharing those keys are sent into the second pass, which compares each name in its original form and returns the names with the highest similarity score.



ID	Name	High Recall Key
68-123	Haugland	HKLNT
68-124	Hagelin	HJLN
68-125	Hoagland	HKLNT
68-126	Heslin	HJLN
68-127	Hoaglin	HKLNT

ID	Name	Associated Data
68-123	Haugland	SSN
68-124	Hagelin	SSN
68-125	Hoagland	SSN
68-126	Heslin	SSN
68-127	Hoaglin	SSN

ID	Name	Associated Data
1.0	Haugland	SSN
.9	Haugland	SSN
.85	Hoaglin	SSN

Matching names are returned to the calling application with a confidence score from 0% to 100%. Setting a minimum match threshold further controls the quality and quantity of results returned.

The Over Watch API may be used to access other information associated with each entry—such as relationships, geographic locations, and pointers to external databases—to help identify specific individuals and places.

Integrated into end-user applications such as anti-money laundering, fraud detection, and watch list comparisons, Over Watch is a lightweight component adding highly accurate name matching without adding heavy memory requirements or slowing the transaction rate.

6.2. Key Features

- Matches names of people, places, and organizations
- Results are ranked by similarity with a confidence score from 0-100%
- Returns partial matches when data is incomplete
- Matches names from many languages, regardless of how the names are written.
- Supports names in: Arabic, Chinese, Dari, English, Farsi, Japanese (hiragana and katakana scripts), Korean (hangul and hanja scripts), Pashto, and Urdu

6.3 Key Benefits

- Speed

- Accuracy (precision and recall)
- Small footprint

7. SUMMARY

Four types of methods are most frequently used to score name similarity:

- **Common key** - These methods, such as Soundex, reduce names to a key or code based on their English pronunciation, such that similar sounding names share the same key. Common key methods are fast and produce high recall (finds most of the correct answers) but have generally low precision (i.e., contain many false hits). Precision is yet lower when matching non-Latin script names, which first must be transliterated to Latin characters to use this method.
- **List method** - This method attempts to list all possible spelling variations of each name component and then uses the name variation lists to look for matches against the target name. The result can be slow performance if very large lists must be searched. Furthermore, this method will not match name variations not appearing in its lists.
- **Edit distance** - **This** approach looks at edit distance, that is, how many character changes it takes to get from one name to another. For example, "Catherine" and "Katherine" have an edit distance of 1 since the "C" is substituted for "K." Edit distance methods work for Latin-to-Latin name comparisons, but precision suffers as each edit is weighted similarly, so a replacement of "c" for "k" is considered equal to a replacement of "z" for "t."
- **Statistical similarity** - A statistical approach trains a model to recognize what two "similar names" look like so that the model can take two names and assign a probability that the two names match or not. This method produces high precision results, but may be slower than the common key method.

7.1. Best Practice: Hybrid Method

Currently, the recommended approach is a hybrid. The hybrid method—where using two or more methods allows the strengths of one to compensate for the weaknesses of the other—proves superior to any single method. An example of a hybrid implementation is Over Watch, which uses a common key "first pass" at matching, taking advantage of the methods' high-recall and speed. A second pass over the pool of candidates resulting from the first pass uses the slower but highly precise statistical method to make up for common key method's low precision.

The first pass gives the system the speed needed for high-transaction environments, and the slower second pass over a small pool of candidate’s re-compares names directly in their original script for greater precision.

8. SUPPLEMENTAL MATERIAL

8.1. Which languages use which scripts?

Arabic, Persian, Pashto and Urdu use the Arabic script with a few language-specific characters added to the Persian, Pashto, and Urdu alphabets. For example, the characters and a- are only used in Persian and not Arabic.

Chinese, Japanese, and Korean use "Han" ideographic characters in their writing, but the specific set of characters and their meanings may vary between the three languages.

Chinese is exclusively written using ideographs (漢字 called *hanzi* in Chinese) where each character represents a concept and has an associated pronunciation. Foreign names are written using characters which approximate the sounds of the foreign name or the meaning of the foreign name

Japanese uses Chinese ideographic characters (漢字 called *kanji* in Japanese) in addition to two syllabic alphabets — *katakana* カタカナ (for words of foreign origin) and *hiragana* ひらがな (for words of Japanese origin and to represent word endings and prefixes). Japanese names are written in a mix of *kanji* and *hiragana*, whereas foreign names are written in *katakana* and approximate the sound of the name.

Korean uses hangul (한글) a phonetic alphabet to write most Korean, reserving Chinese ideographic characters (漢字) called hanja in Korean) for personal names and some scientific terms. Koreans approximate the sound of non-Korean names when transliterating to hangul, similar to Japanese. Hangul enjoys a wider range of sounds than Japanese katakana, but spelling variations may appear, especially for lesser-known names.

8.2. Evaluating Name Matching: Precision vs. Recall vs. F Score

Precision and recall are metrics used to evaluate the quality of search results, whether searching for articles on a topic, or finding name matches.

Suppose you are searching for red balls from a box that contains 7 red balls and 8 green balls. Blindfolded, you pull out 8 balls, of which 4 are red and 4 are green.

Precision is the number of correct items over the total number of items found. That is, you found 4 red (which you wanted) and pulled out 8 balls in total, thus precision is 4/8 (half of your results were correct) or 50%.

Recall asks the question: "Of all the correct items, how many did **I** find?" **In** this case, there were 7 correct items (because there are 7 red balls) and you found 4, thus your recall is 4/7 or 57%.

F-score is a measure that attempts to balance precision and recall and is often called the "accuracy" of a system.

F score = 2*(Precision*Recall) = (Precision+Recall) Thus our F = 8/15 = .53, or 53%

9. EXPLORE FURTHER

For more information or to request an evaluation of Over Watch, please contact David Walter at dwalter@search365.ai, or on +61 402 767 277