Introduction

Name matching plays a pivotal role in many processes, from database deduplication to vetting names against watch lists for fraud prevention, to identifying a person in a database by name alone. The ease of modern travel and communication has increased both the quantity and the complexity of data from countries and cultures around the world. As a result, deciding whether two names represent the same identity is a substantial technical challenge.

These are the issues consistently encountered in multi-cultural name matching:

- **Name variations due to errors**: data-entry errors such as misspellings, transpositions, or name variants introduced due to missing or added characters, accents, hyphenation, and spaces;
- **Natural name variations**: "also known as" (aka) information such as diminutives, nicknames, aliases, and other variants such as name changes due to marriage;
- **Cultural name anomalies and practices**: "von", "bin", and other practices that are applied and recorded inconsistently;
- **Multiple name spelling variations**: transcription of e.g. Arabic and Chinese names.

This white paper explains how WCC’s software platform ELISE meets name matching challenges with a combination of fuzzy matching and analytics.

**Name variations due to errors**

There are several sources of errors that cause name variations. This section describes some common sources of errors and provides examples of how ELISE accounts for these. It is important to note that this paper does not provide an exhaustive list of the algorithms applied, but uses representative algorithms as examples.

**Variations due to data-entry errors ("typos")**

Data-entry mistakes are quite common regardless of the type of data being entered. These errors include: accidentally omitting a letter or number, transposing a letter or number, or accidentally striking the key adjacent to the intended one.

To compensate for typing errors, ELISE uses a variety of algorithms that determine the match between two names. For instance, ELISE recognizes "Kohn Smith" instead of "John Smith" as a likely error, since it knows that the "K" key is next to "J" on the standard keyboard. ELISE automatically compensates for such data-entry errors.

**Fat Fingered Typing Algorithm**

![Fat Fingered Typing Algorithm](image)

*Figure 1: Compensating for keyboard layout typos*
Another error is transposition of characters. This happens when two characters adjacent to each other are switched on entry. WCC employs an algorithm to recognize similarities caused by such variations.

Finally, data-entry errors may arise from transcribing the spoken form of the name. To compensate for this type of error, ELISE uses one or multiple configurable phonetic algorithms. These algorithms convert the name into its phonetic equivalent, allowing ELISE to compare names by pronunciation.

To determine the match percentage between two names, the names are split into their individual components. For each component, its most likely role is determined (e.g. whether it is a first name or a surname).

To compute the similarity between the components with maximum accuracy, several algorithms can be combined depending on the required functionality. Similarity algorithms such as weighted geometric average and cosine similarity are used to handle additional or missing name components or reversals of name components. ELISE can then account for the impact on the overall match probability for each type of difference found.

Database errors and inconsistencies
When searching or merging information from different databases, issues with inconsistent formats, fields, and data quality are common. The source of these problems could be data-entry errors, programming errors, or differences in database design. For example, one database may store names in two separate fields, while another database uses a single field to store the complete name.

ELISE compensates for typical field errors such as swapped or concatenated name fields, as well as for inconsistencies in database design.

Name variations due to punctuation
Another factor that creates inconsistencies in name entry and searching is the use of punctuation marks like hyphens and apostrophes. For example, the name O’Hara may be entered or searched for as O’hara, o’Hara, Ohara, or OHara. Other common entry and search errors include inserting hyphens into two-word last names that do not have a hyphen, or omitting hyphens from names that should contain one. A specialized name compounding and decompounding algorithm compensates for missing punctuation marks and spaces, incorrect punctuation marks, and variations in capitalization.
ELISE applies specific algorithms for each type of error described above, as well as for other errors. A match probability is calculated for each algorithm, and then the results from all algorithms are used to provide a final match score. These results are optimized using statistical analysis, domain knowledge, and historical information. A default configuration is provided that can be used out-of-the-box to start the optimization process.

**Natural name variations**

So far, we have described situations where two names should be treated as equivalent despite data-entry errors. Another class of challenges includes name variations due to natural processes.

- **“Also known as” information**
  
  Many people are known to friends and colleagues under a name variant or nickname. For example, people that were given the name “William” at birth often go by diminutives and variants like “Bill”, “Will”, or even “Liam”. Unlike many systems that normalize such variants on ingest or indexing, ELISE preserves the original data to provide more nuanced results, a particularly important capability when dealing with large data sets.

ELISE ships with a default set of these name variations, which can be extended to fit a particular identity solution. All match settings, including these name variations, can be configured and activated in seconds without the need to reload all data into ELISE. It is also possible to disable the use of these diminutives on request, or to provide specific overrides as required.

When loading data into ELISE, it is possible to specify that for a certain identity, multiple names are known. People who try to mask their identities often use multiple names, pseudonyms, or aliases.
In ELISE, multiple names can be associated with a single identity, ensuring that whatever name is used to search for that identity, all known information about that person can be found.

People may also have multiple names for legitimate, rather than fraudulent, reasons. In Western countries, it is common practice for women to start using their husband’s name upon marriage. The maiden name should, of course, be preserved and used in the matching process. ELISE accommodates all these types of alternate names.

**Handling initials**

It’s also possible that the identity data or query does not contain the full personal name, but just an initial. For example, if the user is registered as “J Doe”, that identity will still come up when searching for “John Doe”. ELISE handles this through its ‘initial expansion and reduction’ functionality. Of course, ELISE searches intelligently, only considers components that are actually designated as personal names.

**Strong transcription capability**

When transcribing a written name between script systems, e.g. from an Arabic script system to a Roman character based system, a phonetic pronunciation is used to create a spelling. This inexact method produces a variety of spellings for a single Arabic name.

One of the most notable examples of inconsistent transcription is the wildly varied spelling of the name of Libya’s former leader. Moammar Qaddafi, Mo’ammar Gadhafi, and Muammar Kaddafi are just a few of over 100 variations. Even major publications like Newsweek, The New York Times, and the Library of Congress each use a unique spelling. To handle such transcription variations, ELISE offers a name

![Figure 5: Incorporating transcription variations](image)
transcription algorithm (see figure 5) and comprehensive
look-up tables containing variations for every language and
script system.

Another example is the Arabic name احمد ابراهيم which
can be written using a phonetic approximation in several
ways, one being Ahmad Ibrahim. Entering the transcription system will render three possible spellings for the first name alone: Ahmad, Ahmed, and Achmed. ELISE will automatically use variations in native script and transcribed variations of both first and last name, as well as other information in the query, to return the best match.

Note that it is possible to search using one script and receive results that are in a different script. In the example above, the query used a name in Arabic script, and received results in Roman script. It is also possible to search using Roman script and receive results in a mix of Roman and Arabic scripts. Of course, in all situations, all relevant name variants in all scripts are taken into account. Transcription is a native capability of the ELISE platform not limited to Arabic names: ELISE can match Chinese and Japanese names as well as many others.

Match result optimization
In addition to detecting human-introduced errors and natural name variations, ELISE incorporates several other mechanisms to improve the match results returned to the user.

» Gender check
Even when comparing just two names, more information can be used than the mere character sequences that form the names. Additional information, such as the most likely gender, can be derived from the names themselves. This information can be used to improve the accuracy of the matches.

For many given names, the person’s gender may be predicted with known probabilities. For example, John is very likely to be a male name, but Joan is very likely to be a female name. Although these two names are similar [an algorithm based on spelling errors alone might rate them a probable match], when the likely gender is included, the probability of a false match is significantly reduced.

Other gender extractions are culture-specific. For instance, many Slavic languages use specific name endings to indicate that the person is female, such as the “–ová” ending in “Suková”.

This cross-property inference not only improves the matching process, but can also be used to flag anomalies in the database. For example, names that are registered incorrectly as male or female can be flagged to prevent embarrassing errors in formal social titles.

» Handling less relevant words
Not all words are equally relevant to a person’s name – for example, titles like “Mr.”, “Sr.” and “Doctor”. These words can cause errors and inconsistency if not handled properly. ELISE has several mechanisms for preventing these errors, such as:

1. Data cleansing to handle words like “Mr.” and “Doctor”, which are not part of a person’s name. Depending on the situation, these words can be flagged to be ignored during matching, or stored in a separate field for e.g. formal professional or personal titles.

2. Term weights to handle culture-specific name prefixes and suffixes such as “Jr.” and “Sr.”. These words are considered to be part of the name, but are less relevant in determining the match between two identities than other words in the name.

3. Sorting with prefixes and suffixes: Prefixes and suffixes can introduce differences in ordering. For instance, common prefixes in the Netherlands are “van”, “van de” or “van der”, like in “Kim van der Wiel”. When sorting a list of names, this name would appear under “W”, probably as “Wiel, Kim, van der”.

4. Cultural significance and variation: Prefixes and suffixes are often culture-dependent. The Dutch prefix “van”, as it appears in “Kim van der Wiel”, should be treated as a less relevant word, but in the Korean name “Kim Van”, “Van” should be treated as a regular family name. In ELISE, this is solved by annotating all tokens with their role in the name, depending on the most likely culture.
**Sorting the results**

After determining the probability that two names match, ELISE ranks all found matches from best to worst. ELISE’s flexible configuration mechanisms allow the user to specify the desired combination of algorithms, and the parameters for each algorithm. This returns the match results precisely in the order in which the user can best process them. Some applications may sort results purely according to match score, while others may group together identities with certain similarities. In contrast, ELISE can sort on any combination of criteria to provide the exact ranking the user requires.

**Other ELISE benefits for name matching**

- **Deterministic and explainable**
  One of the special capabilities of ELISE is the deterministic nature of its score calculation. This has two benefits. First, it explains how a certain match result was achieved. This allows analysts to investigate all the steps involved in matching two names and, if desired, to tweak the match configuration. Second, the match results do not change unexpectedly over time. If the match score between two names is 85% at one point in time, the same match with the same configuration will result in the same score a week later.

- **Application Specific Accuracy**
  Specific match settings can be set up for each application. This can make a single data set available for searching from different applications or processes without the need to load the data multiple times. Thanks to this open architecture and flexibility in configuring the matching engine, ELISE provides optimal matching results for any challenge.

- **Extensible**
  The algorithms provided with ELISE can be augmented with algorithms, rules, and datasets from third-party providers, or ones the customer already has developed. This ensures that ELISE can always be configured and tuned for optimal accuracy.

**WCC’s development team**, in addition to doing its own research and innovation, closely follows the scientific research community to bring the latest innovations to the ELISE matching platform.

**Proven and Best of Breed name matching**

**MITRE challenge**

In the MITRE Multi-cultural Name Matching Challenge, ELISE was recognized as one of the three Top Tier Vendors. The Challenge was to determine the top identity matching technologies as part of MITRE’s ongoing research for the Department of Defense, the Federal Aviation Administration, the Internal Revenue Service and Department of Veterans Affairs, the Department of Homeland Security, and the Administrative Office of the U.S. Courts.

**Examples of ELISE deployment**

- **The Netherlands – immigration service**
  The Dutch Immigration and Naturalization Service, IND, is responsible for registering and processing asylum and visa applications and for issuing residence ID cards to successful applicants. IND uses ELISE’s name matching capabilities to check applicant data against the legacy biographic information in its back-office system. ELISE’s biometric capabilities were also implemented to increase security. The extensibility of the software allowed both biographic and biometric data to be combined in a single search operation.

- **USA – interconnected systems at Concentra**
  Health services company Concentra uses ELISE to centralize patient lookup records from over 200 offices and 85 data centers across the US. When Concentra employees search the database for patient records during admittance, small mistakes such as misspellings may occur. Thanks to WCC’s fuzzy logic technology, ELISE can find entries that are close, but not perfect matches. The same technology validates new patient entries, ensuring that the new patient is not already in the system. In summary, ELISE provides instant patient lookups from this consolidated system while overcoming data-entry errors to accurately identify patients.
EU – Visa Information System
ELISE delivers advanced name matching capabilities for the Visa Information System (VIS) established by the European Commission. Through quick, safe, and secure biometric verifications, VIS delivers faster border checks, more accurate visa procedures, better protection of travelers against identity theft and more security. A consortium comprising Accenture, Morpho and HP was selected to develop and implement this system, and it chose to use ELISE name matching technology because of its excellent performance and functionality.

Beyond Name Matching
ELISE’s capabilities do not end at name matching. Other capabilities of ELISE that can easily be combined with name matching include:

- Use of address and location information for geo-spatial analysis;
- Incorporation of biometric information, such as fingerprint, face, or iris images;
- Enrichment of data, such as the estimated age and gender extracted from facial images;
- Named entity extraction from documents, as well as full text and semantic search in documents.

Figure 6: Returned transcribed variants with match scores
About WCC

WCC Smart Search & Match is the world’s leading supplier of search and match software solutions and services. Founded in 1996, WCC focuses on two specific solution areas: Identity matching and Employment matching. Its ELISE software platform excels in these areas because it searches and matches data in a unique way that yields more meaningful results than any other software. ELISE is designed to search through vast amounts of data from various sources and return relevant results in under a second. It will search and match data in almost any form, using advanced algorithms, contextual knowledge, and other proprietary expertise. The data can be exact or inexact, structured or unstructured, private or public, and combine multiple modalities, both biographic and biometric.

WCC’s long-term experience in developing and supporting employment and identity matching solutions makes it an expert in these fields. WCC understands the business of its customers and knows how to optimize the effectiveness of searching and matching.

WCC’s primary customers are government organizations and large enterprises worldwide. In Identity matching, WCC supplies solutions for border management, justice & public safety, and civil identity. In Employment matching, WCC supplies solutions to public employment services, staffing companies and large enterprises for their corporate HR.

WCC is headquartered in Utrecht, the Netherlands, and also has offices in the USA.