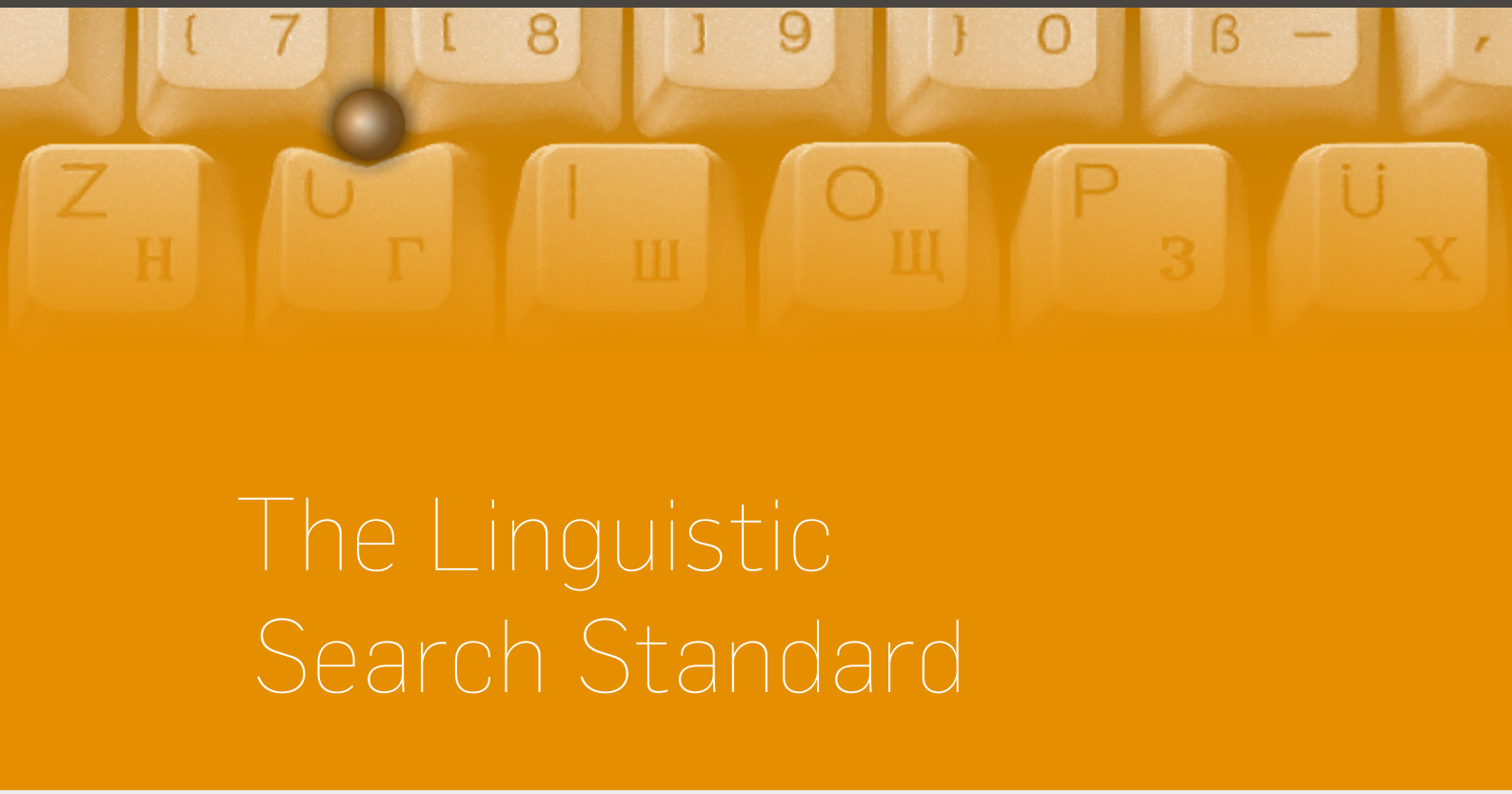
linguistic
search solutions



The Linguistic Search Standard

White Paper on the Guiding Principles for Global Name Matching

The standard has been designed by **Linguistic Search Solutions AG**, in conjunction with leading compliance technology vendors and intelligence data providers as well as industry and regulatory experts. For further information please contact
info@linguisticsearchsolutions.com.
or visit
www.linguisticsearchsolutions.com.

Bringing clarity to the global name matching process

The Linguistic Search Standard for global name matching defines the principle requirements for searching for proper names within international data sets. These principles represent the standard requirements for searching for names presented in a Latin script, regardless of their cultural origin or original language script.

The standard has been formulated in three parts: the principles for determining a near exact match – the Precise Match Level – the additional principles which should be followed to determine very similar matches – the Close Match Level – and the final principles which combine to determine a wider range of matches – the Broad Match Level.

The principles apply to each name element, or part, of the full name. In most cases, name parts are separated from each other by a space in the name, so that the name John Robert Smith contains three distinct name parts. However, principles 1 and 2 require a more flexible approach to be adopted when identifying individual name parts in order to account for the matching of transcription variants and names where individual parts can be merged.

The Precise Match Level

The Precise Match Level defines the requirements for identifying name parts which are essentially the same. There are six guiding principles which set the minimum requirements for meeting the Precise Match Level.

Principle 1:
Different transcriptions of the same names originating from Non-Latin scripts should be considered a Precise Match.

The application of this principle ensures that identical names from Non-Latin script languages will always be matched, provided that a prominent transcription standard has been followed.

Names which originate from a non-Latin script language may be presented differently when transcribed into the Latin script, depending on the transcription standard applied. For example, the Russian name "Ельцин" is commonly presented as "Yeltsin" in English texts, as "Jelzin" in German texts, and as "Eltsine" in French. This principle requires that names originally from non-Latin script languages should be matched if presented in the Latin script using one of the most prominent transcription variants. It also requires that, where a name is presented in Latin characters with diacritics, any loss of these should not prevent a Precise Match from being identified.

For the purposes of this principle, a flexible definition of name parts must be adopted so that transcription variants for names like عبد الرحمن, for example Abd al-Rahman, Abdul Rahman and Abdurrachman, can be identified as such.

There are often very many different transcription standards which can be applied to each original script. This principle requires that variants formed using one of the most prominent transcription standards should be matched. Typically, these will include those used in English, French, German, Italian, Spanish and Portuguese text, though each of these languages may employ several different transcription standards for each script.

There are over 40 non-Latin script based languages that are spoken by communities of over 10 million people. This principle requires that transliterations from each of these major source languages into the 6 target languages listed above are identified as Precise Matches. However, it should be noted that the Standard has been defined to be used with matching systems based on a Latin script. For this reason, it does not require that cross-script matches (e.g. from Cyrillic to Arabic) be identified.

Principle 2:
Names composed of identical name parts should be matched regardless of whether any of the parts have been merged.

The application of this principle ensures that the optional merging or hyphenation of name parts does not prevent a match from being identified.

In some languages, distinct name parts can be presented separately, hyphenated or merged. This is most importantly a consideration when matching names of Eastern origin (for example "Jiangtao", "Jiang-Tao" or "Jiang Tao") or Middle-Eastern origin (such as the Arabic name "Abdal Karim" or "Abd al-Karim" or the Persian name "Alinezhad" or "Ali Nezhad"). However, the merging of given names is also common in some Western cultures, for example the German name "Hanspeter" or "Hans-Peter", and this principle also requires that these Precise Matches should be recognised.

Principle 3:
Names composed of identical name parts should be matched regardless of the order in which the parts are presented unless the order contributes to identification.

The application of this principle ensures that a change in the order in which name parts are presented does not prevent a match from being identified, unless the order of the name parts is culturally or linguistically significant.

The order in which name parts are presented is often altered, particularly when the original name comes from a culture which places the family name first. For example, the Chinese name "Wang Jianhua" should match "Jianhua Wang". However, it should be noted that this principle does not apply where identity is bound to a particular order of name parts, for example in compound Spanish surnames such as "González Lopez" and "Lopez González".

Principle 4:
Identical names from Non-Western backgrounds should be matched regardless of the way in which they have been aligned to a Western name structure.

The application of this principle ensures that Precise Matches cannot be overlooked as a result of data structure and storage practices.

Many global identity data stores have rigid data structures, frequently based on the Western naming convention of first, middle and last names. In many societies, identity data does not readily fit this pattern and names from such cultures may be entered to the data store in more than one way. For example, "Saddam Hussein Al-Tikriti" may be stored in the database with "Hussein" as a middle name, but should still match if "Hussein" is included in the search name as the surname. However, this principle only applies to names from cultures where ambiguity may occur when the names are stored in a Westernised data format. A record with the first name "James" and last name "Martin" should not match a record with first name "Martin" and last name "James" at the Precise level, because Western names are less likely to be erroneously recorded in this way.

Principle 5:
Established nick names and abbreviations should match to their corresponding full name parts.

The application of this principle ensures that the use of common nick names or standard abbreviations of common name parts cannot prevent a Precise Match from being identified.

Many names may be presented in diminutive versions, such as "Bill" for "William", or "Ted" for "Edward", and the abbreviation of common words in the names of groups and organisations, such as "Ltd." or "Corp." is standard practice. The use of such standard diminutives or abbreviations should not prevent a Precise Match from being recognised.

Principle 6:
The omission of peripheral name parts should not prevent a match from being identified.

The application of this principle ensures that the omission of a peripheral name part, such as a title, should not prevent the identification of a Precise Match.

Many data sources may include peripheral name parts such as professional titles and postnominals, such as academic qualifications and generation indicators. In the case of legal entities, these peripheral name parts may include the legal form of the entity, or the geographical location of a branch. This principle requires that the omission of one such name part will not prevent a name from matching.

The Close Match Level

The Close Match Level defines the requirements for identifying name parts which are very similar. The following four principles are additionally required in order to complete the minimum requirements for the Close Match Level.

Principle 7:
Name parts which are both spelt and pronounced in similar ways should be identified as a Close Match.

The application of this principle ensures that similar sounding variants which are also spelt in a similar way are correctly identified as potential matches.

There are many examples of similar names which are hard to tell apart phonetically. These include the forenames "Markus" and "Marcus" or "Steven" and "Stephen", or the family names "Meier" and "Meyer" or "Thomson" and "Thompson". The level of similarity in the pronunciation of different syllables may vary from language to language, so that "Setzer" and "Sezer" sound the same in German, despite their apparent dissimilarity to native English speakers.

Principle 8:
All identical names should be matched regardless of the way they have been parsed for storage.

The application of this principle ensures that names parts being stored as different elements of a Western name should not prevent a Close Match from being detected.

This principle expands on Principle 4 to include names from Western cultures, and other backgrounds where ambiguity in parsing names into a Westernised structure would not normally be expected. The name "James Martin" would, therefore, match the name "Martin James" at the Close Level.

Principle 9:
Common spelling mistakes should not prevent a Close Match from being identified.

The application of this principle ensures that Close Matches are not overlooked as a result of the most common minor spelling mistakes.

In many global name matching processes, data quality issues and the potential for human error can lead to the introduction of minor spelling variations in identity data. The most common errors include the transposition of two characters, the replacement of a character by one that is positioned close to it on a keyboard, or the introduction of an erroneous additional character by hitting an adjacent key on a keyboard. By matching only these more common errors, this principle requires that "Jordan" would match with "Jordam" but not with "Jordas".

However, the Standard does not require that this principle be applied to names of 5 characters or fewer, in order to balance the risk of generating excessive false positive hits.

Principle 10:
The inclusion or omission of less significant name parts should be ignored.

The application of this principle ensures that the omission of less significant name parts should not be considered a mismatch.

This principle requires that the omission of common or otherwise less significant name parts, such as "Der", "Von" or "De La" should be ignored rather than considered as a mismatch. For example, "de Winters" compared with "Winters" should be considered as one matching name element at the Close Level rather than one matched and one mismatched name element.

The Broad Match Level

The Broad Match Level builds on the Close Match Level to include the final principles that allow the additional identification of name parts which match at a broader level.

Principle 11:
Other minor spelling mistakes should not prevent a Broad Match from being identified.

The application of this principle ensures that matches are not overlooked as a result of less common, but still minor, spelling mistakes.

This principle expands on Principle 9 to allow for less common, but still relatively minor spelling mistakes. Under this principle "Capelli" would match with "Capella" on the Broad Match Level, despite the replacement of an "i" with an "a" being a relatively less common spelling mistake. Again, to avoid excessive false-positive matches, this principle is not applied to names consisting of 5 or fewer characters.

Principle 12:
All phonetically similar name parts should match, regardless of the way in which they are spelt.

The application of this principle ensures that names which are pronounced the same way should be matched, regardless of how differently they are spelt.

This principle expands on principle 7 to include matches between phonetically similar names which are spelt in more substantially different ways, such as "Leicester" and "Lester". Again, the phonetic characteristics of each language must be taken into account so that, under this principle, the French name parts "Baudaint" and "Bodin" should be identified as a potential match although they may be pronounced differently by an English speaker.

Principle 13:
All translations of the same name part should match, regardless of their phonetic similarity.

The application of this principle ensures that, where appropriate, names which have been translated from one culture to another will match at the Broad Match Level, even if they have no other similarity.

This principle has been defined to ensure that names which have been translated by individuals moving between cultural regions are still identified as a Broad Match. Under this principle, for example, "Ivan" would match with "John", to account for Eastern Europeans named "Ivan" who might use the name "John" in Western regions. However, the principle does not apply to family names, as Mr Smith should not match M. Lefèvre, Herr Schmied, Sig. Ferrari or Gosp. Kuznetsov. This principle does, however, apply to terms used in the names of legal entities, so that ABC (Deutschland) Ltd. should match ABC (Germany) Ltd.

Calculating the Match Level of a Full Name

The principles described above define the way in which matches between individual name parts should be assessed. In determining the closeness of the match of a name as a whole, the strength of the matches between each individual name part should be taken into consideration. The Standard recommends that the following limits should be set for distinguishing between Precise, Close and Broad matches of full names.

- Precise Full Name Matches
  At least 80% of the component name parts match at the Precise Level.
- Close Full Name Matches
  At least 75% of the component name parts match at the Close or Precise Level; or the name has more than two component name parts, all of which match on at least the Broad Level.
- Broad Full Name Matches
  At least 66% of the component name parts match on at least the Broad Level.

The allowance for the inclusion of additional name parts provides for more flexible matching of full names. It is common that official records may not hold an individual's complete name, particularly where the full name contains many parts. For this reason, it is important to allow for the apparent mismatch of some parts of a multi-part name. For example, "Claire Anne MacDonald" should match at the Broad Level with "Claire Louise MacDonald", to account for an individual whose full name was "Claire Anne Louise MacDonald".

Applying the Linguistic Search Standard

The Linguistic Search Standard is a set of guiding principles. As such it is independent of specific software applications and can be implemented using a number of different technical approaches.

The application of the standard is a comprehensive means of ensuring that relevant identity matches are not overlooked. When implemented correctly, the introduction of excessive low quality matches to search results can be avoided.

The Linguistic Search Standard defines the conditions under which a full name match should be generated. The way in which these matches could be prioritised is not set by the standard, and in practice can be realised in many different ways, depending on the context of the matching process.