

Use phonetic name matching

There are two stages to the matching process that matchIT uses; the key stage and the scoring stage. The first stage creates standardised and phonetic keys based on the input data, which allows potential matches to be identified. The second stage scores each pair of potential matches, using phonetic and fuzzy matching. This parameter governs what sort of phonetic algorithm matchIT uses when generating keys and for scoring.

There are three choices available:-

- **soundIT**

matchIT provides a unique phonetic algorithm for name matching, called soundIT. soundIT takes account of vowel sounds and syllables in the name, and, more importantly, determines the stressed syllable in the word. This means that "Batten" and "Batton" sound the same according to soundIT, as the different letters fall in the unstressed syllable, whilst "Batton" and "Button" sound different, as it is the stressed syllable which differs. Another advantage of soundIT is that it can recognise groups of vowels and consonants that form vowel sounds – thus it can equate "Shaw" and "Shore", "Wight" and "White", "Naughton" and "Norton", and "Leighton" and "Layton" (which are all reasonably common English surnames).

This algorithm was developed with extensive testing on a large table of the most common surnames in the UK. Therefore, it is specifically designed to be used with English names. If a file with mostly non-English names is processed through matchIT, then you may want to try the 'Loose' soundIT or Soundex algorithms instead. For US data we recommend that you use soundIT, because it is proven to work well also with Spanish, German and other names that occur commonly in the US. soundIT has been designed with foreign language versions in mind (i.e. for data collected in countries where foreign languages are spoken). These could quite easily be developed, according to demand. Please contact your supplier if you are interested in this.

Note that the keys that matchIT generates are 'Loose' soundIT keys, where all vowel sounds are equated, together with some consonants, such as 'm' and 'n', 'd' and 't', 's' and 'f'. This is so that potential matches are not missed at the key stage; matchIT uses the 'full' soundIT algorithm at the scoring stage, which will separate out false matches from true matches.

- **Loose soundIT**

This option is effectively the same as the soundIT option, except that matchIT uses the 'Loose' soundIT algorithm as described above at the scoring stage. This is for use mainly with non-English names, on which soundIT works less well, and can miss true matches. This option should not be used on files with mainly English names, as it can potentially lead to more false matches.

- **Soundex**

Soundex is a widely-used algorithm (patented just after the First World War!), which constructs a crude non-phonetic key by keeping the initial letter of the name, then removing all vowels, plus the letters H, W and Y, and translating the remaining letters to numbers. It gives the same number to letters that can be confused e.g. 'm' and 'n' both become 5. It also drops repeated consonants and consecutive letters that give the same number e.g. S and C. It only takes the first four characters of the result, or pads it out with zeroes if it is less than four long. Thus all the common spellings and misspellings of the name "Tootill" equate to the same Soundex key: Tootill, Toothill, Tootil, Tootal, Tootle, Tuthill, Totill are all translated to "T340".

The algorithm that matchIT uses is an enhanced version of Soundex, and is for use mainly with non-English names. This option should not be used on files with mainly English names, as it can lead to false matches e.g. Brady, Beard and Broad get the same Soundex key.

Note that, at the scoring stage, **matchIT** performs name comparisons using data from the NAME field, not from the phonetic keys NAME1, NAME2 and NAME3 – this way it can check for simple typing errors such as "Wilson" and "Wislon" which do not match phonetically.
