# Soundex

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling. The algorithm mainly encodes consonants; a vowel will not be encoded unless it is the first letter.

The Soundex code for a name consists of a letter followed by three numerical digits: the letter is the first letter of the name, and the digits encode the remaining consonants. Consonants at a similar place of articulation share the same digit so, for example, the labial consonants B, F, P, and V are each encoded as the number 1.

The correct value can be found as follows:

1. Retain the first letter of the name and drop all other occurrences of a, e, i, o, u, y, h, w.
2. Replace consonants with digits as follows (after the first letter):
   * b, f, p, v → 1
   * c, g, j, k, q, s, x, z → 2
   * d, t → 3
   * l → 4
   * m, n → 5
   * r → 6
3. If two or more letters with the same number are adjacent in the original name (before step 1), only retain the first letter; also two letters with the same number separated by 'h' or 'w' are coded as a single number, whereas such letters separated by a vowel are coded twice. This rule also applies to the first letter.
4. If you have too few letters in your word that you can't assign three numbers, append with zeros until there are three numbers. If you have more than 3 letters, just retain the first 3 numbers.

Using this algorithm, both "Robert" and "Rupert" return the same string "R163" while "Rubin" yields "R150". "Ashcraft" and "Ashcroft" both yield "A261" and not "A226" (the chars 's' and 'c' in the name would receive a single number of 2 and not 22 since an 'h' lies in between them). "Tymczak" yields "T522" not "T520" (the chars 'z' and 'k' in the name are coded as 2 twice since a vowel lies in between them). "Pfister" yields "P236" not "P123" (the first two letters have the same number and are coded once as 'P').