SfinxBis

Searching for people in web and ldap catalogues is not easy. You have to know the correct spelling of the name. In English you can use Soundex or variants to index names in a standard way and thereby make it easier to search without knowing the exact spelling.

In 1992 two students at Uppsala universitet did an dissertation at basic higher education level regarding swedish name indexing. They called their result Sfinx (“Svensk fonetisk namnindexering”) and it's based on the phonetic algoritm Phonix. Phonix is a Soundex variant.

The goal for SinxBis is to make an implementation of an extended version of the dissertation result available to all SWAMI members. The dissertation is not available on electronic media.

SfinxBis is designed for surnames but works rather well with given names. Every different surname (and given name) for a person shall be coded by itself. For example Jan-Erik Pettersson Olson gives four codes, one each for Jan, Erik, Pettersson and Olson. For effective use of SfinxBis you should save SfinxBis name encoding in your database or ldap server.

Further information as reference implementation, test names and ldap schema for SfinxBis is available at [http://www.swami.se/pub/jsp/polopoly.jsp?d=3483&a=58840](http://www.swami.se/pub/jsp/polopoly.jsp?d=3483&a=58840).

SfinxBis algorithm

The phonetic name encoding of SfinxBis is done by this algorithm.

1) Transform the name into capital letters
2) If the name starts with a nobility prefix remove it:
   a. AF
   b. AV
   c. DA
   d. DE
   e. DE LA
   f. DE LAS
   g. DE LOS
   h. DEL
   i. DEN
   j. DES
   k. DI
   l. DO
   m. DON
   n. DOS
   o. DU
   p. E
3) Remove nearby duplicate letters (in Swedish dubbelteckning) in the beginning of the name.

4) Perform translations to modernize and localize name (the order is important):
   a. "STIERN" → "STJÄRN"
   b. "HIE" → "HJ"
   c. "SIÖ" → "SJÖ"
   d. "SCH" → "SH"
   e. "QU" → "KV"
   f. "IO" → "JO"
   g. "PH" → "F"
   h. Vowel + "Ü" → vowel + "J"
   i. Vowel + "Y" → vowel + "J"
   j. Vowel + "I" → vowel + "J"
   k. "H" + consonant → consonant
   l. "W" → "V"
   m. "Z" → "S"
   n. "À", "Á", "Â", "Ä" → "A"
   o. "Æ" → "Å"
   p. "Ç" → "C"
   q. "È", "É", "Ê", "Ë" → "E"
   r. "Î", "Ï", "I", "Î" → "I"
   s. "Ð" → "ETH"
   t. "Ñ" → "N"
   u. "Ô", "Õ", "Ô", "Ö" → "O"
   v. "Ø" → "Ö"
   w. "Ü", "Û", "Ü" → "U"
   x. "Ü" → "Y"
   y. "Ý" → "Y"
   z. "ß" → "TH"
   aa. "ß" → "SS"
   bb. X → X (the rest of the characters are not transformed to another character)

5) Remove all characters that are not "A" – "Ö"

6) Code the first sound in the name (letters that is behind a plus sign is sound determining and is not part of the sound, vowels are defined in a note at the end):
   a. Vowel → "$"
   b. "DJ", "GJ", "HJ", "LJ" → "J"
c. "G" + soft vowel → "J" + soft vowel

 d. "Q" → "K"

e. "C" + hard vowel → "K" + hard vowel

 f. "C" + consonant → "K" + consonant

g. "X" → "S"

 h. "C" + soft vowel → "S" + soft vowel

 i. "SKJ", "STJ", "SCH", "SH", "KJ", "TJ", "SJ" → "#"

 j. "CH" + vowel, "SK" + soft vowel, "K" + soft vowel → "#" + vowel

 k. X → X (the rest of the characters are not transformed to another character)

7) Divide the name in two parts, the first sound (see above) and the rest.

8) Perform phonetic transformations in rest of the name (not to lose important consonant sounds):

 a. "DT" → "T"

 b. "X" → "KS"

 c. X → X (the rest of the characters are not transformed to another character)

9) Code the rest of the name to a number code (letters that sound approximately the same get the same code, letters that is behind a plus sign is sound determining and is not part of the sound, vowels are defined in a note at the end):

 a. "B", "P" → "1"

 b. "C" + hard vowel, "C" + consonant, "C" as last letter, "K", "G", "Q", "J" → "2"

 c. "D", "T" → "3"

 d. "L" → "4"

 e. "M", "N" → "5"

 f. "R" → "6"

 g. "F", "V" → "7"

 h. "C" + soft vowel, "S", "Z" → "8"

 i. Vowel, "H" → "9"

10) Remove all nearby duplicates are removed from the coded rest.

11) Remove all "9" (vowels and "H") from the coded rest.

12) Put together the initial sound and the coded rest.

Note: In Swedish there are soft and hard vowels, the soft is E, I, Y, Ä and Ö and the hard is A, O, U and Å.