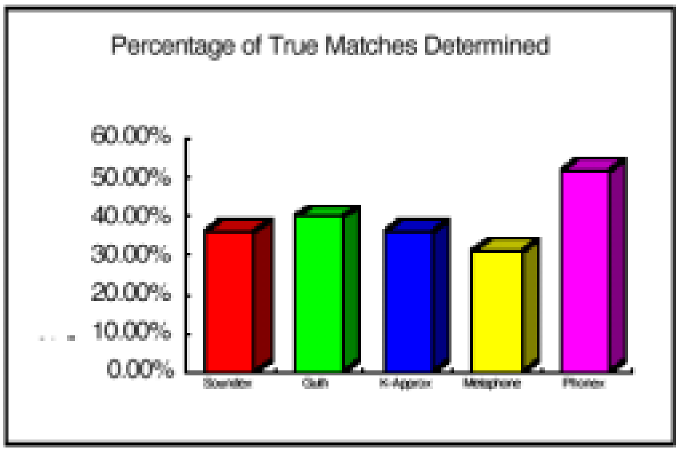
This article was first published in *Computers in Genealogy* Vol. 6, No. 6 (June 1998)

Soundex Update

by Peter Christian

My article on Soundex coding in the previous issue has drawn a number of responses.

First, Brian Randell (*Brian.Randell@newcastle.ac.uk*) has drawn attention to a very interesting article by one of his students. The paper, "An Assessment of Name Matching Algorithms" by AJ Lait and B Randell looks at Soundex and a number of similar algorithms and provides a detailed statistical comparison of their success. Although written from a Computer Science perspective, it is not particularly technical and should be perfectly comprehensible to the non-specialist. Programmers will be interested in the description and code for an algorithm proposed by the author called "Phonex", which seems to offer significant improvements over Soundex. The graph (figure 12b in the original paper) shows the superiority of Phonex, judged by the number of correct matches found.



The paper is available in postscript and ASCII versions at:

http://www.cs.ncl.ac.uk/~brian.randell/home.informal/Genealogy/NameMatching.ps

http://www.cs.ncl.ac.uk/~brian.randell/home.informal/Genealogy/NameMatching.txt

John Hanson (*J_Hanson@compuserve.com*) has been looking at the frequency of some of the letter combinations I discussed in my article and writes as follows:

> The figures that I came up with from the City of London burial index's 17,000+ unique surnames was as follows (All are percentages)

ng	4.66
gh	1.80
mp,mb	1.91
nt,nd	5.79
tch,dg	2.22

> The total is 16.38 percent.
>
> I have generated some new codes based on the above and it resulted in 2,248 new codes. Now comes the problem of attempting to check and see if we can get any better matches.

David Hawgood (*David_Hawgood@compuserve.com*) has written with a useful list of further references:

> Can I draw your attention to some references on Soundex and other name-grouping algorithms from the demography and record linkage disciplines?
>
> Start with the book "Identifying People in the Past" edited by E A Wrigley, 1973, published by Edward Arnold. There are several relevant chapters, including "A brief survey of the algorithmic, mathematical and philosophical literature relevant to historical record linkage", by Ian Winchester - pp. 128-150 of the book.
>
> My other references are in *History and Computing*, the journal of the Association for History and Computing — the current publisher is Edinburgh University Press.
>
> Vol 4 No 1 1992 is a special issue on Record Linkage. The editorial by R J Morris, pages iii to vi, discusses solutions to spelling variations - there are other mentions in papers in the issue.
>
> Vol 8 No 2 1996 includes "Record Linkage Theory and Practice: an experiment in the application of multiple pass linkage algorithms, by Charles Harvey, Edmund M Green and Penelope Corfield, (all of Royal Holloway College at that time), pages 78 - 89. This takes 1784 and 1788 City of Westminster poll books and tries to identify those who voted in both elections. The poll apparently was quite confusing and lively - "the Whig Duchess of Devonshire was said to have kissed a butcher in return for his vote for Fox (she denied it)". Returning to the present, the authors tried 45 different record linkage algorithms and present the results, with a discussion of algorithms which identify matches well, and algorithms which identify all possible matches.

Finally, Barney Tyrwhitt-Drake has produced an updated version of his Soundex program (version 1.3) which can be downloaded from his Web site: *http://www.tdrake.demon.co.uk/*