Martin, for info, Geoff.

A REVIEW OF SOUNDEX NAME MATCHING

How does it work at the moment?

We currently have 3 names databases, two being linked to the same properties database. One is for Rates names only, another combines Housing/GPA names and the third holds Creditor names. The Creditor names database is independent of the properties database.

Each name on the database is assigned a unique 6 character key. Character 1 of the key contains a number in the range 0 to 9. In general if a name contains certain recognised buzzwords, such as LTD or SONS it is classified as an organisation name, and the first character of its key set to 0; otherwise it is classified as a person name, and assigned a key prefix of 1. This 0 or 1 value can be overridden for the purposes of security, and the name assigned a first key character in the range 2-9. Access to that name is then restricted to certain GPA programs and transactions. Names with a first key character greater than 1 are said to be members of the "Restricted Function Set".

Characters 2 - 4 of the key are the 'soundex code', generated from the first few 'significant' characters of the surname; the method of generation is such that H.R. JOHNSON for instance will have the same soundex code as MR.J.K.JAMESON (for more info on how this is achieved see Appendix A). So all people with the same soundex code as JOHNSON will be stored together, forming a 'soundex group'. A soundex group, then, is any group of names with the same first 4 key characters.

A name is submitted (a surname is the minimum requirement) for matching against the database.

The submitted name is passed to a standardisation routine, which calculates its soundex code and assigns 0 or 1 as a first character according to the contents of the name fields.

All names on the database that are within that soundex group are scanned, and matched against the submitted name.

If the submitted name was originally assigned a 0 key prefix, the corresponding soundex group with a 1 prefix is then scanned and matched; or vice-versa: so that the name is matched against both organisations and persons.

The best 99 matches are displayed.

The number of names within a soundex group

Since the advent of Housing, the number of names within certain soundex groups have increased to alarming proportions. Whilst the majority of groups, 90% of organisation groups, and 60% of person groups, have a membership of less than 10, there are some groups which hold a membership of greater than 1000 (see Appendix B).

Large membership soundex groups are the main cause for a degradation of performance of the name matching transactions. Where submitted names belong to such groups, the transactions have to scan through to find the best possible match. It is also extremely likely that the users have cause to access these groups more highly than smaller soundex groups.

If Community Charge involves the merging of Rates, Housing and C.C. into one database, the membership of soundex groups will increase by > 5 fold.

Proposal to use a secondary index to limit the number of names scanned by the name matching transactions

I propose that a secondary index is set up, designed to take as its source and sequencing key a concatenation of the first character of the name's first forename, whether blank or not, and at least the first four characters of the surname key.

The effect of this would be to allow the names to be viewed as being grouped by initial within their soundex group; for instance, submission of 'H JOHNSON' will limit the scan to 'H' names within the JOHN group of names.

The secondary index will only apply to persons, name keys with a first character of '1' to '9'. Soundex groups holding organisations do not have a high membership and it will be difficult to set up organisational names to be referenced by the secondary index because of the current use of buzzwords, i.e. LONDON, used to identify organisations (it could be achieved, but adds complexity).

With the introduction of a secondary index, processing during name matching will be:-

(A) Organisation submitted with no initial:

a) all organisations in that soundex group are scanned as at present

b) the secondary index is used to scan persons with the same name that have no initial.

(B) Organisation submitted with an initial:

a) all organisations in that soundex group are scanned as at present

b) the secondary index is used to scan all persons with the same name that have the same initial.

(C) Person submitted with no initial:

a) the secondary index is used to scan persons with the same name that have no initial.

b) all organisations in that soundex group are scanned as at present

(D) Person submitted with an initial:

a) the secondary index is used to scan persons with the same name that have the same initial

b) all organisations in that soundex group are scanned as at present.

A restriction to the above proposal would be for persons set up on the database with an initial and the user does not have any forename information. The user would be forced to submit a name with initial from A-Z, until found, though if an address is known, property matching is an alternative. Failing this, if the restriction is unpopular, a secondary scan facility could be introduced so that all names on the secondary index, like the name submitted, can be scanned regardless of the initial.

GW/JF

24.6.88

In building the soundex code from the surname fields of a name, the following rules are followed:

a) The following punctuation is ignored: . , / - ( ) '

b) Numbers (in numeric form, such as '435') are ignored.

c) In general, AND, &, and OF are ignored.

d) Where one or two initials appear as the start of the surname and are followed by a longer name (eg. 'I P SHARP' held in the first surname field), those initials will be ignored.

e) The following buzzwords are ignored WHEREVER THEY APPEAR:

THE
LTD
LIMITED
CO
COMPANY
INC
SONS
SON
BROS

f) The following buzzwords are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME:

EXECUTORS
EXORS
BOROUGHS
DEPARTMENT
DEPT
SECRETARY
TREASURER
OFFICER
MANAGER
MINISTRY
MIN
GOVERNOR
GOVERNORS
BURSAR
COMMISSIONERS
CMSRS
MESSRS
MSRS
MSSRS
M/S
MESSR

g) The following phrases are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME:

BOROUGH OF
BORO OF
COUNTY OF
CORPORATION OF
ASSOCIATION OF
ASSOCIATION FOR
CITY OF
UNIVERSITY OF
UNIV OF

h) The following buzzwords are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME, AND THE NAME HAS NOT ALREADY BEEN IDENTIFIED AS THAT OF AN ORGANISATION:

MR
MRS
MISS
MS
MASTER
MESDAMES
MISSES
DR
DOCTOR
COUNCILLOR
CLLR
ALDERMAN
PROF
PROFESSOR
ADMIRAL
BRIGADIER
MAJOR
COLONEL
CAPT
CAPTAIN
REV
SIR

i) AEIOUHWY are ignored.

j) Where two consonants adjacent to each other (or separated by AEIOUHW or Y) are the same, or in the same 'group', the second is ignored. The groups are as follows:

BFPV
CGJKQSXZ
DT
L
MN
R

k) With four exceptions, only the first four 'significant' consonants are used to build the key ('significant' consonants being the ones that remain when all the rules above have been applied). The four exceptions are names beginning with LONDON, NATIONAL, BRITISH, or CLARK (/CLARKE/CLERIC/..); in those cases, five consonants are used.

In building the soundex code from the surname fields of a name, the following rules are followed:

a) The following punctuation is ignored: . , / - ( ) '

b) Numbers (in numeric form, such as '435') are ignored.

c) In general, AND, &, and OF are ignored.

d) Where one or two initials appear as the start of the surname and are followed by a longer name (eg. 'I P SHARP' held in the first surname field), those initials will be ignored.

e) The following buzzwords are ignored WHEREVER THEY APPEAR:

THE
LTD
LIMITED
CO
COMPANY
INC
SONS
SON
BROS

f) The following buzzwords are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME:

EXECUTORS
EXORS
BOROUGHS
DEPARTMENT
DEPT
SECRETARY
TREASURER
OFFICER
MANAGER
MINISTRY
MIN
GOVERNOR
GOVERNORS
BURSAR
COMMISSIONERS
CMSRS
MESSRS
MSRS
MSSRS
M/S
MESSR

g) The following phrases are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME:

BOROUGH OF
BORO OF
COUNTY OF
CORPORATION OF
ASSOCIATION OF
ASSOCIATION FOR
CITY OF
UNIVERSITY OF
UNIV OF

h) The following buzzwords are ignored IF THEY APPEAR AT THE BEGINNING OF THE SURNAME, AND THE NAME HAS NOT ALREADY BEEN IDENTIFIED AS THAT OF AN ORGANISATION:

MR
MRS
MISS
MS
MASTER
MESDAMES
MISSES
DR
DOCTOR
COUNCILLOR
CLLR
ALDERMAN
PROF
PROFESSOR
ADMIRAL
BRIGADIER
MAJOR
COLONEL
CAPT
CAPTAIN
REV
SIR

i) AEIOUHWY are ignored.

j) Where two consonants adjacent to each other (or separated by AEIOUHW or Y) are the same, or in the same 'group', the second is ignored. The groups are as follows:

BFPV
CGJKQSXZ
DT
L
MN
R

k) With four exceptions, only the first four 'significant' consonants are used to build the key ('significant' consonants being the ones that remain when all the rules above have been applied). The four exceptions are names beginning with LONDON, NATIONAL, BRITISH, or CLARK (/CLARKE/CLERIC/..); in those cases, five consonants are used.

LONDON BOROUGH OF TOWER HAMLETS

SCAN GROUPS WITH MEMBERSHIP HIGHER THAN
THE PREDEFINED LIMIT = 000800

30 JUN 88

SOUNDEX GROUP		TOTAL NO. MEMBERS	NO. ORG. MEMBERS	NO. PER. MEMBERS
XA	C200	3410	0	3410
XB	EA40	834	1	833
XM	9000	2294	0	2294
XS	D640	1120	0	1120
XU	BA80	949	1	948
XU	C400	816	0	816

APPENDIX B