Specification of the UID algorithm

The following method works well for tracking the vast majority of individuals across an anonymous data collection but I'm sure you can imagine realistic scenarios where it may fail. This situation can only be resolved with extensive leg work chasing up the individual records, validating the entries and then amending the duplicates to reconcile the variable. If you do not have the resources to perform this level of validation, then the anomalous records can either be excluded case-wise from your analysis, or accepted as a minor source of error. Without allowing for the human breeding season, there is less than a 0.3% chance of sharing a birthday with any given individual while sharing names is obviously far more variable and culturally dependant.

Constructing the statistical linkage key

There are any number of ways this can be done but the method that I have used (with a high degree of success) employs a concatemer of four demographic variables:

§ The 2nd, 3rd and 5th letters of the individual's surname;

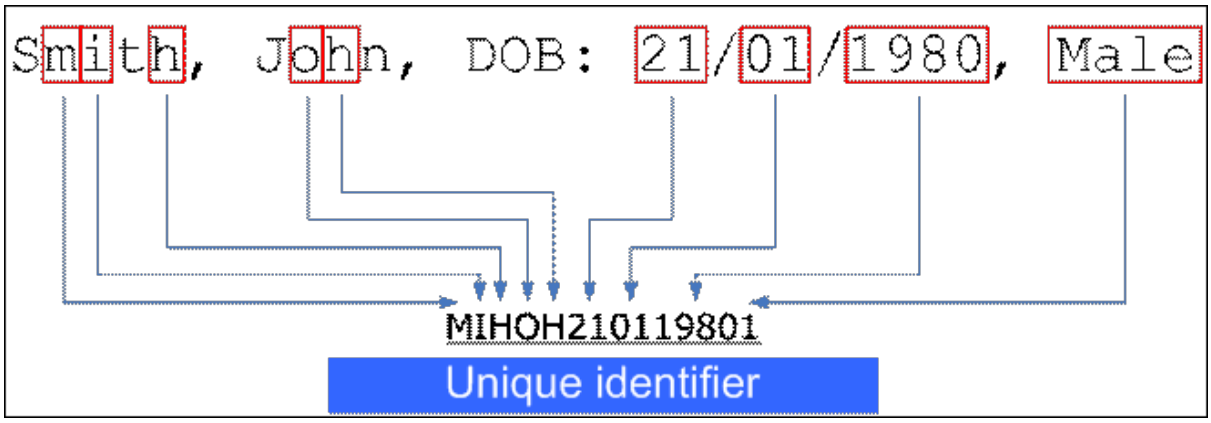
§ The 2nd and 3rd letters of their first name;

§ Their entire birth date without punctuation *; and

§ The person's gender encoded as a 1 (male), 2 (female), or 3 (undisclosed)

* This is more of a technical consideration as many data systems dislike dashes, dots and the solidus

This is summarised in the diagram below:



Additional considerations

Obviously given the great variety of names available, particularly from an international perspective, you'll need some fairly robust rules to govern how people record the information. Here are some of the rules I use:

- Do not include apostrophes, hyphens, inflections, dashes or spaces.
- If the first name or surname of the person is not long enough to supply the requested letters i.e. a surname of less than five letters) then the number '2' should be substituted to reflect the missing letters. The placement of a number '2' should always correspond to the same space that the missing letter would have within the field.
- If the first name or surname of the person is completely absent, it should be replaced by a string of digits of value '9' to indicate 'not stated'. The use of 'not stated' for this data item should be strongly discouraged but such cases can be easily identified and excluded from any analysis as the alphabetical characters have been replaced with numeric ones.

Additional considerations (continued)

- Often people use a variety of names, including legal names, married/maiden names, nicknames, assumed names, traditional names etc. Even small differences in recording, such as the difference between MacDonald and McDonald, can make record linkage impossible. To minimise discrepancies in the recording and reporting of name information, recorders should ask for a person's full 'surname'. I imagine that this shouldn't be a major issue with Australia given the near ubiquitous use of Medicare data.

- In some cultures it is traditional to state the surname first. To overcome discrepancies in recording/reporting that may arise as a result of this practice, recorders should always ask the person to specify their given name and their surname separately.

Examples

'Surname' response

	1st	2nd	3rd	4th	5th	6th	7th
PANG, Ho		A	N		2		
O'DOYLE, Mary		D	O		L		
De VERES, Phil		E	V		R		
MacMILLS, Jo		A	C		I		
Brian		9	9		9		

'Given name' response

	1st	2nd	3rd	4th	5th	6th	7th
PANG, Ho		O	2				
O'DOYLE, Mary		A	R				
De VERES, Phil		H	I				
NIKOV, Steve		T	E				
BEHLER		9	9				