Appendix 2 – Some key anonymisation techniques

Data masking

This involves stripping out obvious personal identifiers such as names from a piece of information, to create a data set in which no person identifiers are present.

Variants:

- **Partial data removal** – results in data where some personal identifiers, eg name and address have been removed but others such as dates of birth, remain.
- **Data quarantining** - The technique of only supplying data to a recipient who is unlikely or unable to have access to the other data needed to facilitate re-identification. It can involve disclosing unique personal identifiers – eg reference numbers – but not the 'key' needed to link these to particular individuals.

These are relatively high risk techniques because the anonymised data still exists in an individual-level form. Electoral roll data, for example, could be used to reintroduce names that have been removed to the dataset fairly easily. However, this type of data is also relatively 'rich' in terms of allowing an individual to be tracked as part of a longitudinal study for example.

Pseudonymisation

De-identifying data so that a coded reference or pseudonym is attached to a record to allow the data to be associated with a particular individual without the individual being identified.

Deterministic modification is a similar technique. 'Deterministic' here means that the same original value is always replaced by the same modified value. This means that if multiple data records are linked, in the sense that the same name (or address, or phone number, for example) occurs in all those records, the corresponding records in the modified data set will also be linked in the same way. This facilitates certain types of data analysis.

This is also a relatively high risk technique, with similar strengths and weaknesses to data masking.

Aggregation

Data is displayed as totals, so no data relating to or identifying any individual is shown. Small numbers in totals are often suppressed through 'blurring' or by being omitted altogether.

Variants:

- Cell suppression - if data is from a sample survey then it may be inappropriate to release tabular outputs with cells which contain small numbers of individuals, say below 30. This is because the sampling error on such cell estimates would typically be too large to make the estimates useful for statistical purposes. In this case, suppression of cells with small numbers for quality purposes acts in tandem with suppression for disclosure purposes.
- Inference Control – Some cell values (eg small ones such as 1-5) in statistical data can present a greater risk of re-identification. Depending on the circumstances, small numbers can either be suppressed, or the values manipulated (as in Barnardisation). If a large number of cells are affected, the level of aggregation could be changed. For example, the data could be linked to wider geographical areas or age-bands could be widened.
- Perturbation – such as Barnardisation - is a method of disclosure control for tables or counts. It involves randomly adding or subtracting 1 from certain cells in the table. This is a form of perturbation.
- Rounding – rounding a figure up or down to disguise precise statistics. For example if one table may have a cell with value of 10,000 for all people doing some activity up to the present date. However, the following month, the figure in that cell rises to 10,001. If an intruder compares the tables it would be easy to deduce a cell of 1. Rounding would prevent this.
- Sampling - in some cases, when very large numbers of records are available, it can be adequate for statistical purposes to release a sample of records, selected through some stated randomized procedure. By not releasing specific details of the sample, data holders can minimise the risk of re-identification.

- Synthetic data - mixing up the elements of a dataset – or creating new values based on the original data - so that all of the overall totals and values of the set are preserved but do not relate to any particular individual.
- Tabular reporting – a means of producing tabular (aggregated) data, which protects against re-identification.
- These are relatively low risk techniques because it will generally be difficult to find anything out about a particular individual by using aggregated data. This data cannot support individual-level research but can be sufficient to analyse social trends on a regional basis, for example.

Derived data items and banding

Derived data is a set of values that reflect the character of the source data, but which hide the exact original values. This is usually done by using banding techniques to produce coarser-grained descriptions of values than in the source dataset eg replacing dates of birth by ages or years, addresses by areas of residence or wards, using partial postcodes or rounding exact figures so they appear in a normalised form.

Again, this is a relatively low-risk technique because the banding techniques make data-matching more difficult or impossible. The resulting data can be relatively rich because it can facilitate individual-level research but presents relatively low re-identification risk.