Anonymisation of qualitative datasets:
Notes on methodology

William Clayton
Centre for Transport and Society
University of the West of England, Bristol, UK

Anonymising data to protect the identity and privacy of participants is a task familiar to most researchers, and it is an issue which will only gain in importance as the acceleration and proliferation of computing technology provides the ability to indefinitely preserve greater and greater amounts of information, and to make this information more easily available than ever before.

Anonymity is generally a central clause in the establishment of informed consent for research participants. With quantitative data, anonymisation may often be a relatively straightforward task, as participants' personal information and identifying characteristics can be transformed into a set of statistics, which retain much of their meaning, but lose their individual identity. With qualitative data however, this task is more challenging, as often a greater depth of personal information and context is recorded, and much of the meaning in qualitative research is derived from its subjectivity. Therefore the difficult balance to be struck in anonymising qualitative data is one between preserving the context and subjectivity of the data, whilst making the participant themselves unidentifiable.

There are several different types of identifier found in qualitative data which individually or collectively could lead to identification of participants. The main types of identifier are listed below:

- Names (of participants, family members, friends, colleagues, acquaintances)
- Place names (of own/family's/friends'/colleagues'/acquaintances': residences, local streets, local shops/amenities, schools/colleges/universities, landmarks, estates, workplaces)
- Personal information (age, nationality, home village/town/city, previous addresses)
- Other idiosyncratic details

The UK Data Archive (2013) explains that whilst it is necessary to find and anonymise these identifiers, nonetheless this task should be carried out in a delicate manner to retain as much of the meaning in the qualitative data as possible. This involves replacement in place of deletion where possible, and a reasoned approach to achieving a robust (but not counterproductive) level of anonymity:

> *"When anonymising qualitative material, such as transcribed interviews, identifiers should not be crudely removed or aggregated, as this can distort the data or even make them unusable. Instead pseudonyms, replacement terms or vaguer descriptors should be used. The objective should be to achieve a reasonable level of anonymisation, avoiding unrealistic or overly harsh editing, whilst maintaining maximum content"*

Below there is an outline of the approach taken in anonymising 36 qualitative interviews exploring cycling behaviour in several locations around the UK. The outline addresses each of the different identifiers above, and discusses other issues which were encountered in the process.

Note: In all cases of removal or replacement, it was standard to denote a changed term using square brackets (e.g. "I said to [husband] that he should cycle more…"). Maintaining this practice ensures that it is always clear exactly where changes have been made.

Names (of participants, family members, friends, colleagues, acquaintances)

In this case, names that could identify individuals were removed and replaced with a descriptor.

- For family members this was possible through using [husband], [wife], [son], [grandmother], etc… In other cases it was appropriate to use terms such as [male friend], [female colleague], etc…
- This form of anonymisation presented few problems, however in the rare instances that it was not possible to ascertain who was being discussed; a more generic descriptor such as [adult male] was used instead.

Place names (of own/family's/friends'/colleagues'/acquaintances': residences, local streets, local shops/amenities, schools/colleges/universities, local landmarks, estates, workplaces)

Greater discretion was necessary in making judgements about the appropriate level of anonymity for place names. The research was exploring cycling behaviour, and in this context it was important to retain as much geographic detail as possible to understand the use of specific cycle routes.

Note: Place names were routinely left intact when referred to by participants *in general terms*. As an example, if a participant specifically discussed themselves or others working at or attending a hospital in their hometown, the hospital's information would be removed; however if they more generally discussed the same hospital as a big local employer, or as a cause of town-centre traffic congestion at rush hour, its details would be left in.

- The basic area of anonymity worked to for areas of residence was the neighbourhood level. This constituted suburbs in cities and towns, and whole villages in the case of rural areas. The names of these and names of the features within them (streets, shops, estates, landmarks, etc…) were replaced with appropriate descriptors.
- Large place names were almost always left in (for example: London, Chester, or Cambridge), with the exception being when these were sufficiently unique to provide a potential identifier (as an example: one participant from a small village community outside Cambridge had previously lived in an unusual international city, and this detail was removed as it could have potentially made them identifiable to someone with local knowledge of this community)
- Workplaces and employer names were replaced with [workplace/employer] as standard, sometimes with additional detail as appropriate when the additional context was useful (for example: [healthcare provider])
- Local schools and colleges which participants or other named people had attended were replaced with descriptors. University names were sometimes replaced and sometimes left in, dependent on context. (for example: if a participant from a suburb of Chester had a child that attended university in Manchester, this detail would be left in; however if a participant from a small village near Cambridge had a child that attended an international university this detail would be removed)

Personal information (age, nationality, home village/town/city, previous addresses)

Personal information including the age, nationality, and previous addresses of participants and other people mentioned were routinely replaced with an appropriate descriptor.

- In the case of age, individuals' ages would be replaced with an age range; the type of range used was varied by context. (For example: for children it was often more useful to put [primary school age], or [mid-teens]. For adults it was more useful to use a [35-45] format)
- Nationality was removed or replaced dependent on context. As with above examples, this was done with consideration paid to the uniqueness of both a participant's nationality and a

participant's location (for example: a person from Scotland living in a suburb of Bristol would not have their nationality anonymised, but a person from Nicaragua living in a small village outside Chester would do). The same approach was adopted for previous addresses and home towns/villages/cities.

Other idiosyncratic details

This is the category in which contextuality played the greatest part, and discretion and a reasoned approach were essential. Idiosyncratic details encountered in this case included unusual medical conditions, unique combinations of household vehicles (for example three classic cars, a camper van, and a tandem bicycle), and unusual hobbies and pastimes.

These details were dealt with on a case-by-case basis. For example in one instance it was necessary to remove the details of a medical condition suffered by a participant as it was rare enough to warrant it being changed to [medical condition]. In other cases, participants discussed family members who had suffered (for example) heart attacks, and in such cases where the condition was relatively common this detail was left intact.

The idiosyncratic details offered by participants in qualitative data should always be treated robustly, and it is best to 'err on the side of caution' in many cases. This is because such details – whilst perhaps relatively innocuous information *by themselves* – hold the potential to be identifiers in conjunction with other information. Indeed, this is true to a lesser degree of all of the personal information provided by a participant about both themselves and others, and it is important that the anonymisation process is conducted with an awareness of how one piece of information which is left intact might form part of an identifying feature in conjunction with another piece (or several other pieces of information). Whilst it is desirable to leave as much context intact as possible, attention must be paid to the potential interactions of different pieces of information.

A useful approach is for the researcher to put themselves in the place of the uninformed reader and attempt to trace the participant using the information that they have left intact. It is useful to have a map available and access to the internet, so that additional context about the data being provided by the participant can be sourced, and a more informed decision on individual cases of anonymisation can be made.

The processes described above will hopefully be of use to anyone seeking to anonymise qualitative data. It should be noted that the list above and the types of personal information discussed in this document are not intended to be exhaustive; the task of anonymising qualitative data will always involve a degree of discretion and problem-solving to ensure that participants' privacy and identity are protected, whilst as much as possible of the important context and subjectivity in the data is retained.

Further useful sources of information on data anonymisation:

UK Data Archive
http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation?index=2

University of the West of England: Data management guidance
http://www1.uwe.ac.uk/library/usingthelibrary/researchers/manageresearchdata/managingresearchdata/guidance.aspx

The Research Ethics Guidebook
http://www.ethicsguidebook.ac.uk/Anonymising-your-data-309

Anonymising Research Data: (Clark, A.) University of Leeds
http://eprints.ncrm.ac.uk/480/1/0706_anonymising_research_data.pdf

Forum: Qualitative Social Research
http://www.qualitative-research.net/index.php/fqs/article/view/511/1102