

Anonymisation techniques and other measures to enable using and sharing research data

Managing and Sharing Research Data workshop
London, 2 December 2009

Using and sharing confidential research data

...obtained from people as participants

Requires a combination of:

- discussing consent and confidentiality with participants / respondents (dialogue)
- anonymisation of data
- user access regulations

researchers only; use license with confidentiality agreement; data unavailable for certain time period

What is required depends on:

- nature of research
- planned data uses
- is study specific

Identity disclosure

A person's identity can be disclosed through:

- direct identifiers

name, address, postcode, telephone number, voice, picture

usually NOT essential research information (administrative)

- indirect identifiers – possible disclosure in combination with other information

occupation, geography, unique or exceptional values (outliers) or characteristics

Why anonymise data?

- Ethical reasons
 - protect people's identity (sensitive, illegal, confidential info)
 - disguise research location
- Legal reasons
 - not disclose personal data (DPA)
- Commercial reasons

Essential points

- Never disclose personal data (unless specific consent)
- Reasonable / appropriate level of anonymity
- Maintain maximum meaningful info
- Where possible replace rather than remove
- Identifying info may provide context, do not over-anonymise
- Re-users of data have the same legal and ethical obligation to NOT disclose confidential info as primary users

Anonymising quantitative data

- Remove direct identifiers
names, address, institution
- Reduce the precision / detail of a variable through aggregation
postcode sector vs full postcode, birth year vs date of birth, occupational categories
- Generalise meaning of detailed text variable
occupational expertise
- Restrict upper / lower ranges of a variable to hide outliers
income, age

Relational data

Extra care needed - combinations of related datasets or a dataset in combination with publicly available info can disclose information

e.g. businesses studied are mapped in publication

Geo-referenced data

Spatial references (point coordinates, small areas) may disclose position of individuals, organisations, businesses, etc.

Removing spatial references prevents disclosure; but all geographical, locational and related information lost

Alternatives:

- Keep spatial references intact and impose access restrictions on data instead
- Reduce precision - replace point co-ordinates with larger, non-disclosing geographical areas
km² area, postcode district, ward, road
- Reduce precision - replace point coordinate with meaningful variable typifying the geographical position
catchment area, poverty index, population density

Anonymising qualitative data

- Plan or apply editing at time of transcription or initial write up
- Except: longitudinal studies - anonymise when data collection complete (linkages)
- Avoid blanking out information
- Use pseudonyms or redactions
- Removing or aggregating unusable, unreliable or unrepresentative data
- Consistency within responses
- Identify replacements
- Create anonymisation log of changes made; store this log separately
- XML mark-up can be used

Example: Anonymisation log interview transcripts

| Interview / Page | Original | Changed to |
|------------------|-----------------------|------------|
| Int1 | | |
| p1 | Spain | European |
| p1 | E-print Ltd | Printing |
| p2 | 20 th June | June |
| p2 | Amy | Moira |
| Int2 | | |
| p1 | Francis | my friend |
| | | |
| | | |

`<seg type="anonymised"`

Audio-visual data

Digital manipulation of audio and image files can remove personal identifiers

voice alteration, image blurring (e.g. of faces)

Labour intensive, expensive, may damage research potential of data

Alternatives:

- Obtain consent to use and share data unaltered for research purposes
- Avoid mentioning disclosing information during audio recordings

Tips

- Always consider anonymisation of research data together with consent agreements and access restrictions
- Regulating / restricting user access may offer a better solution than anonymising
- Avoid collecting data that need anonymisation
 - do not ask for full names if they latter need to be removed from data*
- Remove, mask, change identifiers
- Maintain maximum information
- Retain unedited versions of data for use within the research team and for preservation
- Plan at start of research, not at the end

Sources

- Clark, A. 2006. Anonymising research data. NCRM Working Paper Series 7/06. ESRC National Centre for Research Methods.
[http://www.ncrm.ac.uk/research/outputs/publications/WorkingPapers/2006/0706_anonymising_research_data.pdf]
- Inter-University Consortium for Political and Social Research (ICPSR). 2005. Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle. 3rd Edition. ICPSR, Ann Arbor.
- UK Data Archive 2008. Manage and Share Data - Anonymising research data
[<http://www.data-archive.ac.uk/sharing/anonymise.asp>]
- UK Data Archive 2009. Managing and Sharing Data – a best practice guide for researchers.
[<http://www.data-archive.ac.uk/news/publications/managingsharing.pdf>]
- Timescapes meetings & discussions

Exercises / scenarios

- Anonymising qualitative data:
 - Foot and mouth study Cumbria 2001-2003 (5407)
 - Conflicts and violence in prison (4596)
- Anonymising quantitative data: Labour Force Survey
- Confidential relational and geo-referenced data: British Household Panel Survey