

Data anonymisation and access control to data

Data Management and Sharing Workshop
Vienna, 14-15 April 2010 WISDOM

Why anonymise research data?

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

When using or sharing research data

...obtained from people as participants

Always consider jointly:

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

E.g. researchers only; use licence with confidentiality agreement; data unavailable for certain time period

Identity disclosure

A person's identity can be disclosed through:

- direct identifiers

E.g. name, address, postcode, telephone number, voice, picture

often NOT essential research information
(administrative)

- indirect identifiers – possible disclosure in combination with other information

E.g. occupation, geography, unique or exceptional values (outliers) or characteristics

Key points for anonymising

- Never disclose personal data - unless consent for disclosure
- 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
E.g. names, address, institution, photo
- Reduce the precision/detail of a variable through aggregation
E.g. birth year vs. date of birth, occupational categories, area rather than village
- Generalise meaning of detailed text variable
E.g. occupational expertise
- Restrict upper lower ranges of a variable to hide outliers
E.g. income, age
- Combining variables
E.g. creating non-disclosive rural/urban variable from place variables

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

Remove spatial references - prevents disclosure; also all geographical and related information lost

Better:

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

Anonymising qualitative data

- Plan or apply editing at time of transcription
 - Except: longitudinal studies - anonymise when data collection complete (linkages)

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

- Avoid blanking out; use appropriate replacement text
- Avoid over-anonymising text can distort data, misleading
- Consistency within replacement text
- Identify replacement text
- Keep anonymisation removals made – keep record
- XML mark-up can be used for anonymisation

`<seg type="anonymised">word to be anonymised</seg>`

Audio-visual data

Digital manipulation of audio and image files can remove personal identifiers

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

Labour intensive, expensive, may damage research potential of data

Better:

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

What if anonymising is impossible?

- Obtain consent for sharing non-anonymised data
- Regulating/restricting user access:
 - archived data usually not in public domain
 - use of data for specific purposes only (e.g. research)
 - restrict who can use data , e.g. approved researchers only
 - data access authorisation from data owner needed
 - embargo for given time period
 - secure access to data
- Researchers - consider access to data and safe storage

Conclusion

- Always consider anonymisation of research data together with consent agreements and access restrictions
- Regulating/restricting user access may offer better solution than anonymising
- Avoid collecting data that need anonymisation
 - E.g. do not ask for full names if they later need to be removed from data*
- Remove, mask, change direct identifiers
- Maintain maximum information
- Retain unedited versions of data for preservation
- Plan anonymising 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 2009. Managing and Sharing Data – a best practice guide for researchers.
[<http://www.data-archive.ac.uk/news/publications/managingsharing.pdf>]
- <http://www.data-archive.ac.uk/sharing/anonymise.asp> and
<http://www.data-archive.ac.uk/sharing/accessrestriction.asp>

Exercises/scenarios

- Anonymising qualitative data
Foot and mouth study Cumbria 2001-2003 (SN5407)
- Anonymising quantitative data
Polish and Lithuanian migrant workers (SN6284)
- Confidential relational and geo-referenced data
British Household Panel Survey