ANONYMISATION

LIBBY BISHOP
RESEARCH DATA MANAGEMENT TEAM
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

LOOKING AFTER AND MANAGING YOUR RESEARCH DATA
UK DATA ARCHIVE, COLCHESTER, 24-25 APRIL 2013

UK • DATA ARCHIVE

WHY ANONYMISE RESEARCH DATA?

Ethical reasons

- protect people's identity (sensitive, illegal, confidential info)
- disguise research location

Legal reasons

- not disclose personal data (DPA)

Commercial reasons

Discuss with your research participants

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

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

ANONYMISING QUALITATIVE DATA

- plan or apply editing at time of transcription

  except: longitudinal studies - anonymise when data collection complete (linkages)

- avoid blanking out; use pseudonyms or replacements

- avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading

- consistency within research team and throughout project.

- Identify replacements, e.g. with [brackets]

- keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files

GEO-REFERENCED DATA

Spatial references (point co-ordinates, 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. km^2 area, postcode district, ward, road
- reduce precision - replace point co-ordinate with meaningful variable typifying the geographical position; or summary statistics of location
  e.g. catchment area, poverty index, population density

ANONYMISING QUALITATIVE DATA

- not collect disclosive data unless necessary
- plan or apply editing at time of transcription

  except: longitudinal studies - anonymise when data collection complete (linkages)
- avoid blanking out; use pseudonyms or replacements
- avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading
- consistency within research team and throughout project
- identify replacements, e.g. with [brackets]
- keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files
- xml mark-up can be used for anonymisation

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

ANONYMISING QUALITATIVE DATA

Example: Anonymisation log interview transcripts		
Interview / Page	Original	Changed to
Int1		
p1	Spain	European
p1	E-print Ltd	Printing
p2	20th June	June
p2	Amy	Moira
Int2		
p1	Francis	my friend

P31. Joan → Mary
P97. Carol → {Mother}
P34. Colchester → {Town in S.E.England}
P65. Welshpool High School → @@##High School##@@

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, e.g. at UK Data Archive:
  - archived data NOT in public domain
  - use of data for specific purposes only after user registration
  - data users sign legally binding End User Licence – e.g. not identify any potentially identifiable individuals
  - stricter access regulations for confidential data (case by case basis):
    - access to approved researchers only
    - requiring data access authorisation from data owner prior to data release
    - confidential data under 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

CONTACT

UK DATA ARCHIVE
UNIVERSIY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX CO4 3SQ

T +44 (0)1206 872143
E: datasharing@data-archive.ac.uk
W: www.data-archive.ac.uk