De-identification and pseudonymisation of qualitative data

Veerle Van den Eynden
UK Data Service
University of Essex

Managing and sharing research data: best practice for data protection

London, 28-29 November 2018

UK Data Service

Scenario Gender in Society Perceptions Study

- Household, community and public level data and information on key risk factors for gender inequality and threats for violence affecting women and girls in the Kyrgyz Republic
- Research on issues of *(1)* women's political participation, *(2)* women's economic empowerment, *(3)* violence against women in the form of bride kidnapping and child marriages, *(4)* women's religious radicalization and its implications for migration and *(5)* women's involvement in labor migration.
  - Focus group discussions
  - Interviews
  - Case studies
  - Census of incidents (child marriage, bride kidnapping)

Scenario - protect anonymity measures

- Not record any official identifying data
- Let participants choose a pseudonym for interview
- Password protect interviewee contact details
- Do not connect in any way pseudonyms to the password protected interviewee contact details

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*

Disclosure review and anonymisation

- Before sharing data, check whether people can be identified from the data
- Check consent – if researching people, have they agreed to have their information shared for research?
- Disclosure review – generally a two-stage process, then de-identification where needed - some automation possible

Disclosure review stage 1: direct identifiers

Direct identifiers directly identify participants:

- Are there any direct identifiers in your data?
- Unless explicit consent obtained from the participants for public sharing, direct identifiers should always be removed from data; for qualitative data disguise as pseudonym

Disclosure review stage 2: indirect identifiers

- Key identifying information: age/birth date, education, employment; religious affiliation, geographic area
- Check combinations of information, e.g. employment + small geography
- Sensitive information: health/medical; crime; drug/alcohol use etc.

Always consider risk vs. utility of anonymised data

De-identifying textual data

- Plan or apply editing at time of transcription

  except: longitudinal studies – de-identify 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
- Show replacements, e.g. with [brackets]
- Keep a log of all replacements, aggregations or removals made – keep separate from de-identified data files
- Text anonymisation helper tool can help you find disclosive information to remove or pseudonymise in text files
  - MS Word macro to find and highlight numbers and words starting with capital letters in text, which are often disclosive, e.g. as names, companies, birth dates, addresses, educational institutions and countries

Example anonymisation

Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer
Date of birth: 2 May 1965
Gender: Male
Occupation: Frontline worker
Location: Plumpton, North Cumbria

Lucas was living at home with his parents, "but I'm hoping to move out soon" so we met at his parents' small neat house. We sat in a very comfortable sitting room with an open fire and Lucas made me coffee and offered shortbread. Although at first Lucas seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget abut the tape.

I will just start by asking you to tell me a little bit about yourself and your background.

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Comment [v1]: Replace: Ken
Comment [v2]: delete
Comment [v3]: delete
Comment [v4]: Replace: Ken
Comment [v5]: Replace: Ken
Comment [v6]: Replace: Ken

UK Data Service

What if anonymising is impossible?

- Obtain consent for sharing non-anonymised data
- Regulate or restrict user access

Exercise: qualitative data

Seymour, Jane (2010-2012). Managing suffering at the end of life: a study of continuous deep sedation until death. [Data Collection]. Colchester, Essex: Economic and Social Research Council. 10.5255/UKDA-SN-850749

Questions

Veerle Van den Eynden

veerle@essex.ac.uk