How to anonymise qualitative and quantitative data

Maureen Haaker

Anca Vlad

16 April 2021

Overview

- Introduction: Why anonymise?
- A very short introduction to anonymisation theory
- Anonymisation or Pseudonymisation?
- Access restrictions
- Exercise
- 4 steps to anonymisation
- A couple considerations for qualitative data
- Exercises: de-identification
- Further resources and questions

What is disclosure? Why do we need anonymisation?

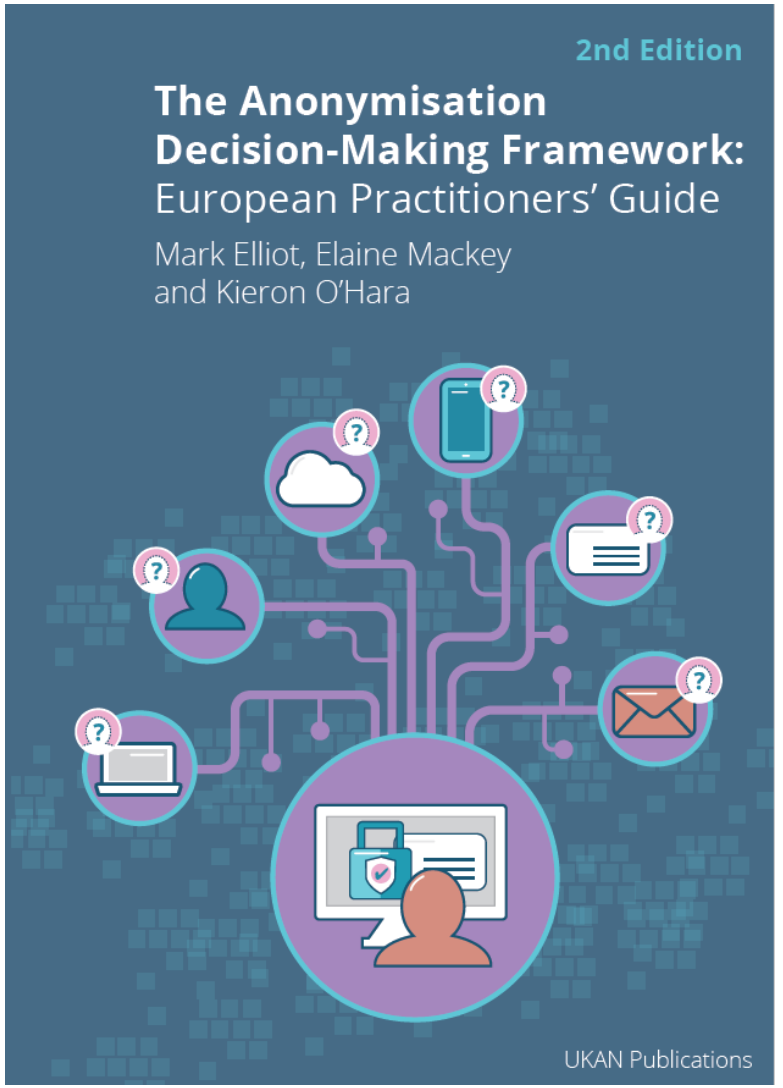
- ✓ Disclosure = identification
- ✓ Disclosure happens when someone is able to identify a data subject from data or information they have access to from one source or multiple sources.
- ✓ Different types of disclosure: identity, attribute, inferential
- ✓ Anonymisation is a process that attempts to prevent disclosure or identification of data subjects from a specific dataset
- ✓ Anonymisation and pseudonymisation is part of Statistical Disclosure Control (SDC): the aim of SDC is to minimise/mitigate the risk of identification to an acceptable level that still allows researchers to maximise data use (use the data to it's full potential or as close to it as possible)
- ✓ When disclosure risk  information loss 

Anonymisation theory

Data situation audit

Risk analysis and control

Impact management



Legal obligations,
or when you need to break confidentiality

How will the data be used?

I'm asking for permission to use anonymised quotations and narrative themes, along with any photographs and video you provide in the interviews or diaries for research purposes. All diaries and interviews from all participants will be analysed together for common themes about what everyday life is like when pregnant. As I work through my analysis, I will transcribe any audio recordings or handwritten diary entries. As I transcribe, I'll anonymise any identifying details, such as your name and address. All digital files will be saved on a password-protected computer at University of Essex and all paper documents will be stored in a locked drawer at my office at the University of Essex, to which only I have access.

Throughout the project, I will be the only one with access to un-anonymised data, and my supervisors will have access to anonymised data. Since this project has gone through ethical approval from the Health Research Authority, NHS Trust staff may also be audit this project to ensure I am protecting your information appropriately, and may ask to see relevant sections of data.

I may need to break this confidentiality if you disclose illegal or criminal activity to me or I become aware of an issue that puts you or a child's safety at risk. In this instance, I will aim to first discuss the issue with you, but I may be legally obliged to share this information with the appropriate authorities.

Anonymisation / Pseudonymisation

Anonymised data:

- '...information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." (Recital 26, GDPR)
- Cannot re-identify data subjects (even the data owner)

Pseudonymised data:

- “…the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.” (Article 4, GDPR)
- Identifiable data has been removed or redacted so that cannot be traced back to the real values. Re-identification of data can only be achieved with knowledge of the de-identification key or by combination.

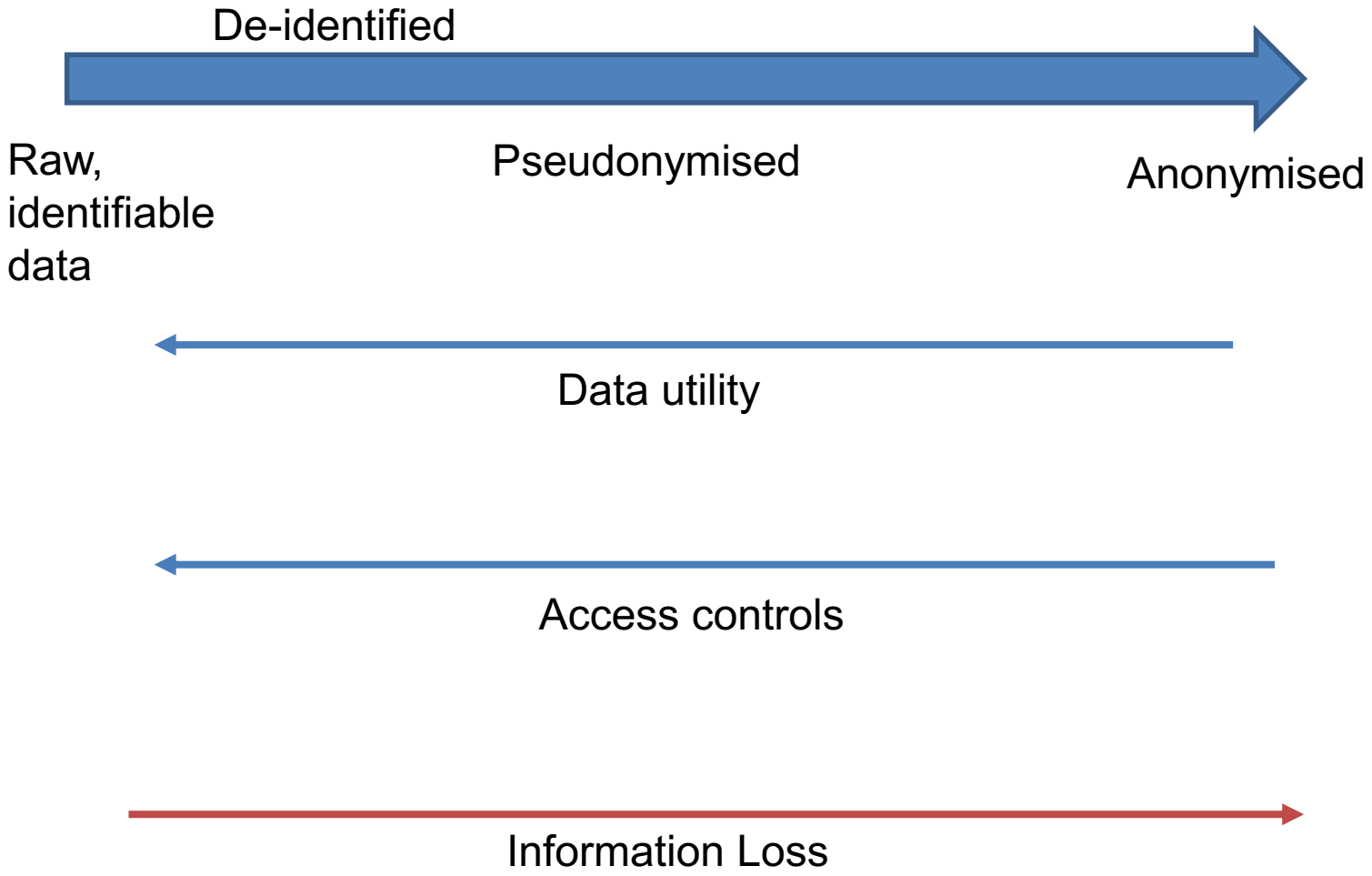
Anonymisation / Pseudonymisation

- ICO: ‘re-identification’: describes the process of turning anonymised data back into personal data through the use of data matching or similar techniques.

- The DPA does not prohibit the disclosure of personal data, but any disclosure has to be fair, lawful and in compliance with data protection principles.

- To consider:
  - ✓ the age of the information (less sensitive over time, but consider ethical)
  - ✓ level of detail
  - ✓ context: private life or about more public matters, such as their working life, or life satisfaction?
  - ✓ Rule of thumb: try to assess the effect – if any - that the disclosure would have on any individual concerned

Anonymisation / Pseudonymisation



Data governance, managing access to data

Open

- available for download/online access under open licence without any registration

Safeguarded

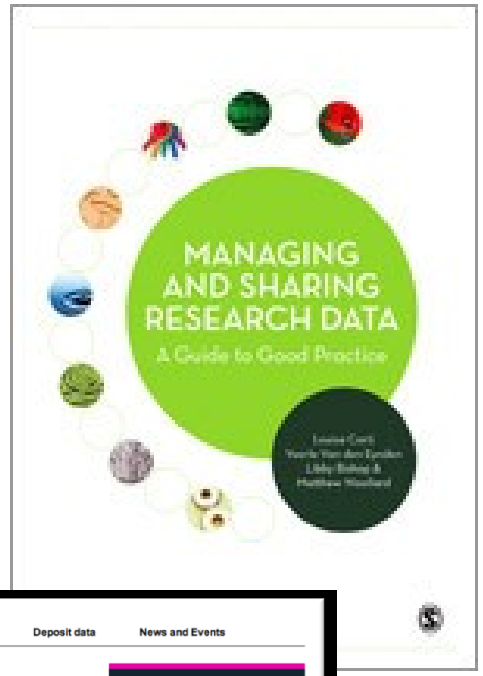
- available for download / online access to logged-in users who have registered and agreed to an End User Licence (*eg. not identify any potentially identifiable individuals*)
- special agreements (depositor permission; approved researcher)
- embargo for fixed time period

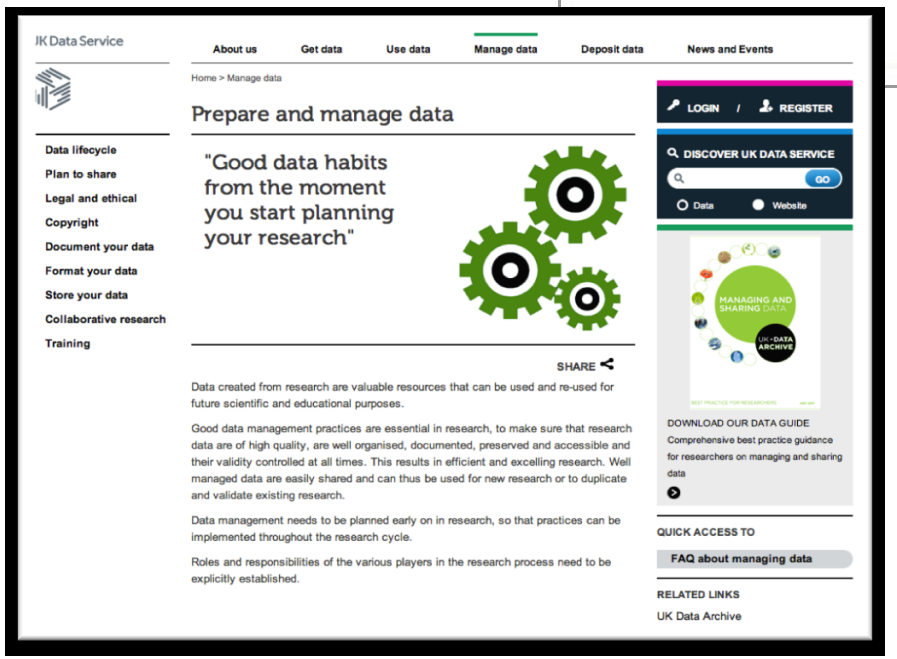
Controlled

- available for remote or safe room access to authorised and authenticated users whose research proposal has been and who have received training

UKDS data management guidance

- Best practice guidance: www.ukdataservice.ac.uk/manage-data.aspx
- Managing and Sharing Research Data – a Guide to Good Practice:(Sage Publications Ltd)
- Training: www.ukdataservice.ac.uk/news-and-events/events
- Twitter: @UKDSRDM





Classifying information (variables)

A. Identifying variables

1. Direct identifiers - information that *directly* identifies data subjects
   - examples: social insurance number, names, address, national insurance number, IP address etc.
2. Key identifiers - information that in combination, may uniquely identify data subjects;
   - can potentially be linked to other sources of data as well (such as the electoral register)
   - examples: gender, age, region, occupation, income

B. Sensitive variables - information that is often subject to legal and ethical concerns;
- examples: criminal history, sexual preferences and behaviour, political affiliations, medical records, income
- can lead to attribute disclosure even if identity disclosure is prevented.

One variable can be both identifying and sensitive. Example: income.

You are not so anonymous!

Anonymisation, Step 1

Similar for both quantitative and qualitative data, the first step is always to identify and remove or redact identifying information (direct identifiers).

- Easier for quantitative data - removal of variables
- Can vary for qualitative data
    - replace with pseudonyms or not redact out

Anonymisation Quantitative Data (Step 2)

- Identify all indirect identifiers:
  - ✓ Age/Date of Birth
  - ✓ Gender
  - ✓ Occupation
  - ✓ Income
  - ✓ Geography (area/county/city/village etc.)
  - ✓ Ethnic Background/Ethnicity
  - ✓ Religion

- Note here how important good quality metadata can be for this process (variable labels, value labels).

Anonymisation Quantitative Data (Step 3)

- Look at frequencies in the data to identify potentially disclosive information.
- Look at outliers.
- Look at string variables (other open text) to identify if they contain any personal, potentially disclosive or sensitive information (“I worked for X company for 30 years” or “my brother has a rare type of disease” or “I was a victim of domestic abuse and I used charity x for support”)
- Introduction to Statistical Disclosure Review

Anonymising quantitative data: some tips

- Aggregate or reduce the precision;
- Recode categorical key variables into fewer categories (k-anonymity)
- Suppressing specific values of key variables for some units (k-anonymity)
- Generalise meaning of text variables - replace potentially disclosive free-text responses with more general text
- Restrict the upper or lower ranges of a continues variable to hide outliers
    - E.g age – recode into 70+
    - How to decide? Look at distribution of that variable.

- Anonymise geo-referenced data - replacing point coordinates with non-disclose variables

Useful software

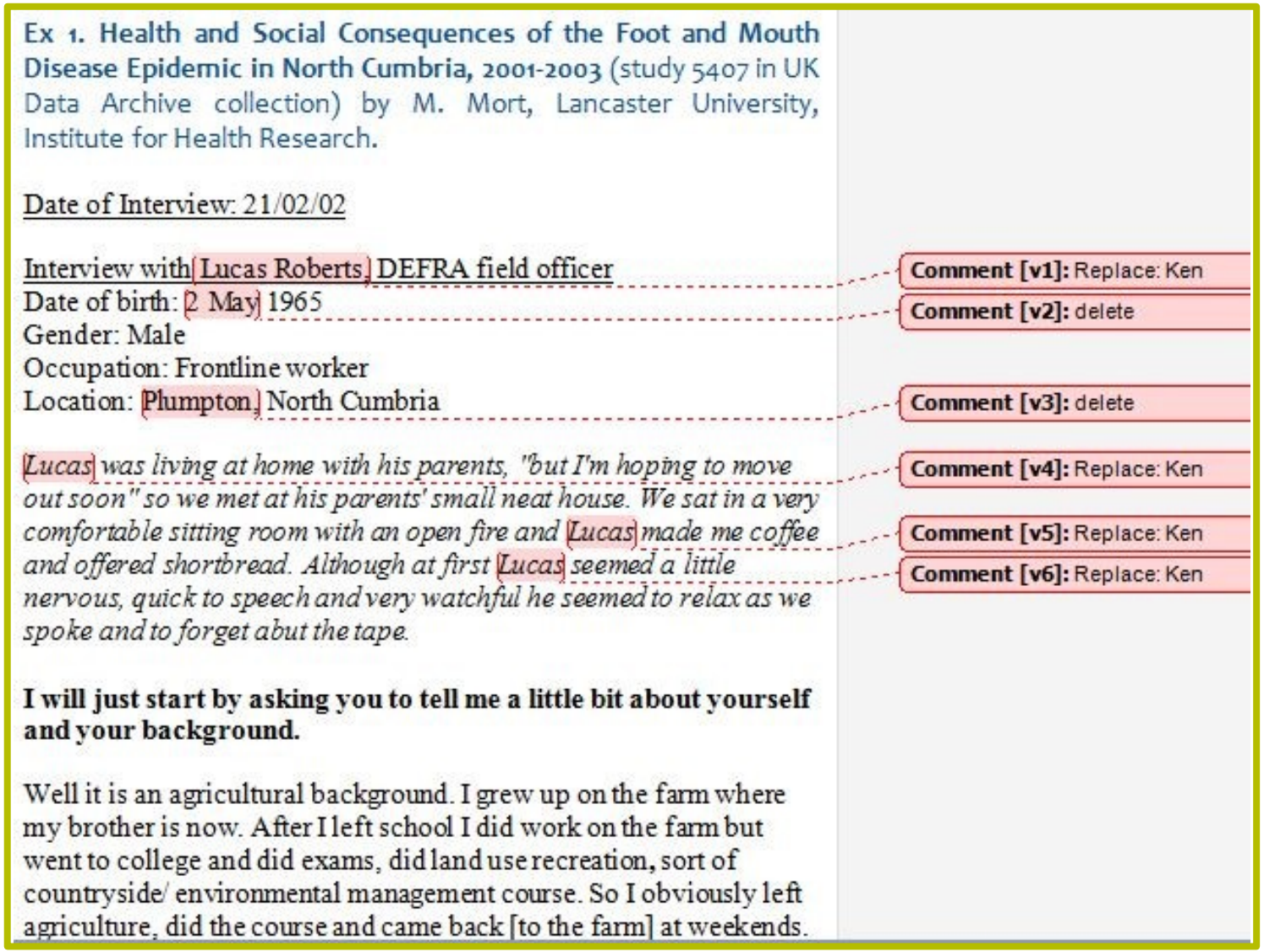
- sdcMicro – R package (free) – has a user friendly interface so minimal coding skills needed.
- QAMyData - UK Data Service developed a free (GitHub) easy-to-use open source tool, that provides a health check for numeric data. The tool uses automated methods to detect and report on some of the most common problems in survey or numeric data, such as missingness, duplication, outliers and direct identifiers.
- ARX - a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analysing the usefulness of output data.
- µ-Argus – developed by Statistics Netherlands; User Manual
- Text anonymization helper tool: Tool to help find disclosive information in textual files. The tool does not anonymize or make changes to data, but uses MS Word macros to find and highlight numbers and words starting with capital letters in text.

Anonymising qualitative data: some tips

- Plan or apply editing at time of transcription
  Except: longitudinal studies - (linkages)
- 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
- Avoid blanking out; use pseudonyms or replacements
- Avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading

Controlling access a better option than over-anonymising

In practice: 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

In practice:
wording in consent forms / information sheets

We expect to use your contributed information in various outputs, including a report and content for a website. Extracts of interviews and some photographs may both be used. We will get your permission before using a quote from you or a photograph of you.

After the project has ended, we intend to archive the interviews at …. Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

How will the data be used?

I'm asking for permission to use anonymised quotations and narrative themes, along with any photographs and video you provide in the interviews or diaries for research purposes. All diaries and interviews from all participants will be analysed together for common themes about what everyday life is like when pregnant. As I work through my analysis, I will transcribe any audio recordings or handwritten diary entries. As I transcribe, I'll anonymise any identifying details, such as your name and address. All digital files will be saved on a password-protected computer at University of Essex and all paper documents will be stored in a locked drawer at my office at the University of Essex, to which only I have access.

Throughout the project, I will be the only one with access to un-anonymised data, and my supervisors will have access to anonymised data. Since this project has gone through ethical approval from the Health Research Authority, NHS Trust staff may also be audit this project to ensure I am protecting your information appropriately, and may ask to see relevant sections of data.

In practice: data with access conditions

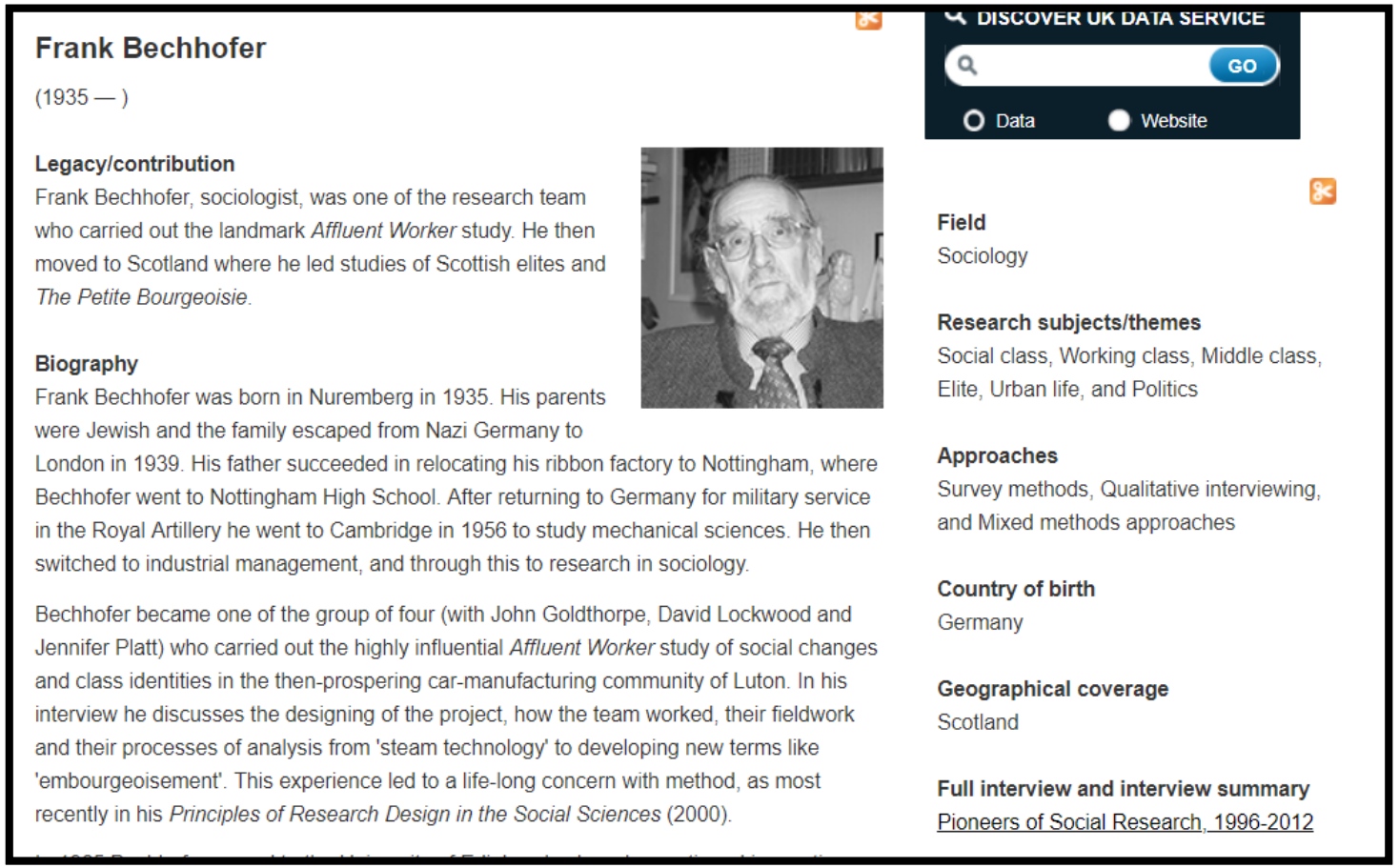
- Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.
- Interviews (audio and transcript) and written diaries with 54 people
    - 40 interview and diary transcripts are archived and available for re-use by registered users
    - 3 interviews and 5 diaries were embargoed until 2015
    - audio files archived and only available by permission from researchers

discover.ukdataservice.ac.uk/catalogue/?sn=5407

doc.ukdataservice.ac.uk/doc/5407/mrdoc/pdf/q5407userguide.pdf

In practice: Pioneers of Social Research

Conducted by pioneering oral historian, Paul Thompson and his colleagues, this collection contains 43 life story interviews with pioneering social researchers, covering family and social background and key influences with detailed accounts of major projects.



Frank Bechhofer

(1935 —)

Legacy/contribution
Frank Bechhofer, sociologist, was one of the research team who carried out the landmark *Affluent Worker* study. He then moved to Scotland where he led studies of Scottish elites and *The Petite Bourgeoisie*.

Biography
Frank Bechhofer was born in Nuremberg in 1935. His parents were Jewish and the family escaped from Nazi Germany to London in 1939. His father succeeded in relocating his ribbon factory to Nottingham, where Bechhofer went to Nottingham High School. After returning to Germany for military service in the Royal Artillery he went to Cambridge in 1956 to study mechanical sciences. He then switched to industrial management, and through this to research in sociology.

Bechhofer became one of the group of four (with John Goldthorpe, David Lockwood and Jennifer Platt) who carried out the highly influential *Affluent Worker* study of social changes and class identities in the then-prospering car-manufacturing community of Luton. In his interview he discusses the designing of the project, how the team worked, their fieldwork and their processes of analysis from 'steam technology' to developing new terms like 'embourgeoisement'. This experience led to a life-long concern with method, as most recently in his *Principles of Research Design in the Social Sciences* (2000).

DISCOVER UK DATA SERVICE
GO
Data Website

Field
Sociology

Research subjects/themes
Social class, Working class, Middle class, Elite, Urban life, and Politics

Approaches
Survey methods, Qualitative interviewing, and Mixed methods approaches

Country of birth
Germany

Geographical coverage
Scotland

Full interview and interview summary
Pioneers of Social Research, 1996-2012

In practice: Managing Suffering at the End of Life

Some dy
symptom
life that c
conventi
circumst
to induce
This prac
deep sed
'palliative
data des
with refra
sensitive
symptom
hastening death.

> **Sensitive personal data** means personal data consisting of information as to -
>
> (a) the racial or ethnic origin of the data subject,
>
> (b) his political opinions,
>
> (c) his religious beliefs or other beliefs of a similar nature,
>
> (d) whether he is a member of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992),
>
> (e) his physical or mental health or condition,
>
> (f) his sexual life,
>
> (g) the commission or alleged commission by him of any offence, or
>
> (h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.

In practice: anonymisation plans

- Project background
- File management
- Mandatory anonymisation
  - Direct Identifiers (names, contact details)
  - Places
  - Ages and dates
- Possible anonymisation
  - Medical information about others not taking part in study
  - Sensitive information (unfavourable opinions of others, details of legal cases, etc.)

3-prong approach to protecting participants: Consent, anonymisation, and access

- Ask for consent to share –researchers must be informed about risks ***and*** benefits of data sharing
- Anonymise – only if damage to data is minimal (not images)
- Regulate access
  - End User Agreement (UK Data Archive)
  - Embargo
  - For selected sensitive or disclosive data – registered users; permission from data depositor

These strategies enable most data to be shared.

Exercise 1

De-identification of quantitative data:

https://ukdataservice.ac.uk/media/622178/exercise_de-identify_quant_data.pdf

Exercise 2

De-identification of qualitative data:

https://ukdataservice.ac.uk/media/622177/exercise_de-identify_quali_data.pdf

Tools and templates

- Model consent form: http://www.dataarchive.ac.uk/media/112638/ukdamodelconsent.pdf
- Survey consent statement: http://dataarchive.ac.uk/media/147338/ukdasurveyconsent.doc
- Transcription template: http://dataarchive.ac.uk/media/136055/ukdamodeltranscript.pdf
- Transcription instructions: http://dataarchive.ac.uk/media/285633/ukda-example-transcriptioninstructions.pdf
- Transcription confidentiality agreement: http://dataarchive.ac.uk/media/285636/ukda-transcriber-confidentialityagreement.pdf
- Data list template: http://dataarchive.ac.uk/media/2989/UK%20Data%20Archive%20Example%20Data%20List.pdf

Further resources

- Anonymising Research Data - ESRC National Centre for Research Methods, Working Paper 7/06
- Guide to Social Science Preparation and Archiving from the Inter-University Consortium for Political and Social Research
- Anonymisation and Social Research, Ruth Geraghty
- Timescapes anonymisation guidelines, University of Leeds
- Anonymisation: managing data protection risk - ICO code of practice
- The Anonymisation Decision-Making Framework - Mark Elliot, Elaine Mackey Kieron O'Hara and Caroline Tudor
- Jisc guidance on anonymous data
- Advice from med.data.edu on anonymisation

Upcoming events

Webinars and workshops

- 22 April, 11-12:30: Getting started with Secondary Analysis (online)
- 29 April, 11-12:00: Data Management Basics 1: Introduction to data management and sharing
- 30 April, 10-11: Data Management Basics 2: Ethical and legal issues in data sharing
- 18 May, 11-12:30: Dissertation projects: introduction to secondary analysis for qualitative and quantitative data
- 27 May, 10-11: Consent issues in data sharing

Other training

- 27 April, 6-7:30 pm: PyDataMCR – Data FAQs with the UK Data Service
- 28 April: Safe Researcher Training (online)
- 19-20 May: Introduction to Understanding Society using Stat/SPSS/R/SAS (University of Essex)
- 21 May: Panel data econometrics using Understanding Society (University of Essex)
- 12-16 July: Essex Summer School

Get connected

http://ukdataservice.ac.uk/about-us/contact.aspx

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKdataservice

https://twitter.com/UKDataService
@UKDataService

https://www.youtube.com/user/UKDATASERVICE

.

Check out our Twitter for more updates.

Questions

Enquiries / Help Desk:

https://ukdataservice.ac.uk/help

help@ukdataservice.ac.uk