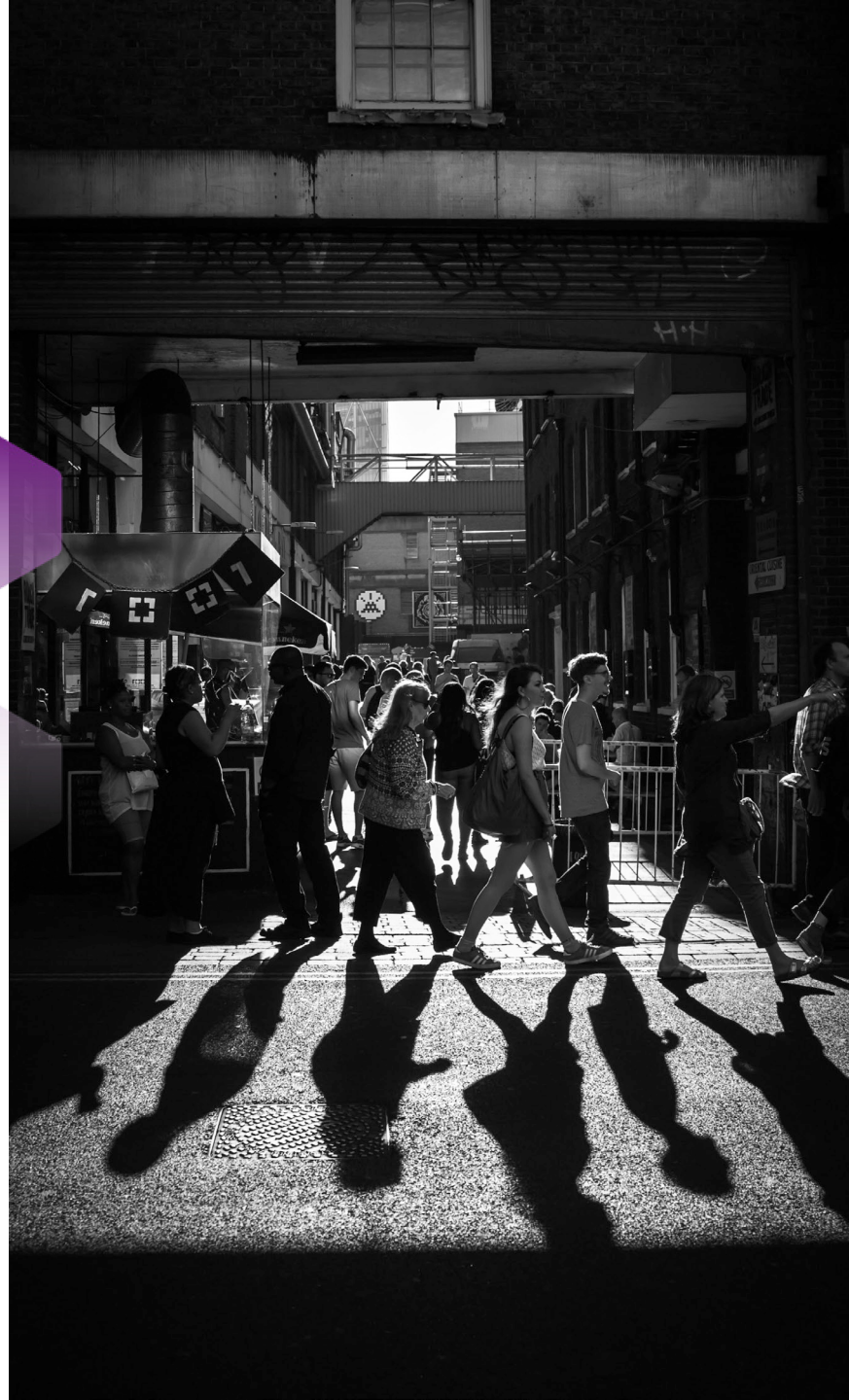
UK Data Service

Introduction to anonymisation techniques for social sciences research data

Cristina Magder
Data Collections Development Manager
Maureen Haaker
Senior Research Data Officer: Qualitative Data
18 November 2024



UKRI Economic and Social Research Council

Format of the workshop

- Presentation
  - A structured overview of anonymisation techniques, practical examples, and key considerations.
- Mentimeter
  - Engage with practical exercises and scenarios to reinforce concepts. Share your thoughts, ideas, and responses in real-time.
- Padlet for Q&A
  - Conclude with a live Q&A session during the workshop, a static version of the Padlet will be made available as an additional resource for the workshop.

Today's programme

- Anonymisation and the data context
- Three-prong approach for social science primary research
- Applicable legislation and wider considerations
- Types of identifiers
- Effective anonymisation and practical considerations
- Common indirect identifiers
- Anonymisation steps
- Anonymisation techniques and considerations for quantitative data
- Anonymisation techniques and considerations for qualitative data
- Q&A session

Today’s main focus

- Types of identifiers: direct identifiers and indirect identifiers.
- Types of data: survey data, transcript data, and audio and visual data

What to look out for and why? Which techniques and options are available?

Before that however let’s look at key data context definitions and the context in which data exists.

Data anonymisations and key considerations

Anonymisation is a valuable tool that allows data to be shared, whilst protecting research participants.

Every data source and every data context is different! A number of key factors to consider are:

- Types of identifiers: direct identifiers and indirect identifiers.
- Types of data: survey data, audio and visual data, transcript data etc.
- Ethical considerations: informed consent, confidentiality and withdrawal rights.
- Data sharing strategies: access control, user agreements and data security.

Three-prong approach to protecting participants

- **Consent:** participants must be informed about risks and benefits of *any* data sharing.
- **Anonymisation:** treat the data reducing the risk of re-identification recognising data utility will be reduced, therefore a balanced approach must be taken
- **Access:**
  - Who? How? For how long?
  - Access levels and user agreements
  - Leverage legislation and existing frameworks such as the Five Safes Framework

These strategies enable most data to be shared.

Data protection considerations

Researchers must adhere to data protection requirements when managing and/or sharing personal data.

In the UK, the Information Commissioner's Office (ICO) provides in-depth information about personal data, including definitions and considerations for anonymisation, within the scope of the UK General Data Protection Regulation and the Data Protection Act 2018.

Personal data

Personal data is defined by the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018. In essence, personal data is information that relates to an identified or identifiable natural person, be it directly or indirectly, taking into account other information derived from published sources.

Special category data

Special category data, also sometimes referred to as sensitive personal data, are defined in Article 9 (1) of the UK GDPR. These data require additional attention and protection due to their sensitivity. Special category data consists of data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data to uniquely identify a natural person, data concerning health, a person's sex life or sexual orientation.

Personal information

Personal information is defined by the Statistics and Registration Service Act 2007 as information that relates to and identifies an individual (including a body corporate). Personal information, compared to personal data, takes into consideration corporate bodies.

Consent and participant communication

It is important to differentiate between consent from a legal view and consent from an ethical view.

Consent can be used as one of the legal basis for processing personal data. However, the most common basis for research conducted with personal data in the UK are:

- Public task (for all public bodies/authorities (i.e. organisations that process data while carrying out tasks in the public interest for example NHS / HSC, Universities, UKRI, etc).
- Legitimate interest (for all non-public bodies for example charities, commercial companies, etc).

Informed consent from an ethical perspective must always be considered no matter what legal basis for processing data is applied. Always **inform your participants** about your plans **for collecting, using and sharing data**.

Access levels: UK Data Service example

Access Options

OPEN

Suitable for data that are not classified as personal data or personal information and with no residual risk of disclosure or where consent to share personal data as collected is in place.

SAFEGUARDED

Suitable for data that are not classified as personal data or personal information, and where the risk of identification is considered sufficiently remote; also referred to as effectively anonymised data as per ICO guidance.

CONTROLLED

Suitable for data classified as personal data or personal information and data that are particularly sensitive, commercially or otherwise. Access is facilitated through the Five Safes Framework.

Types of identifiers

Based on the **personal data** and **personal information** definitions an individual or a body corporate can be identified either **directly** or **indirectly**, hence identifiers are classified as either direct or indirect.

Direct identifiers

Information that directly identifies data subjects.

Indirect identifiers

Information that in combination, may uniquely identify data subjects. It can potentially be linked to other sources of data (such as the electoral register).

Types of identifiers

Direct identifiers

Information that directly identifies data subjects.
Examples: name, address, National Insurance number, NHS number, IP address, email address

Indirect identifiers

Information that in combination, may uniquely identify data subjects. It can potentially be linked to other sources of data (such as the electoral register).
Examples: sex, age, region, occupation, income, ethnicity, religious affiliation

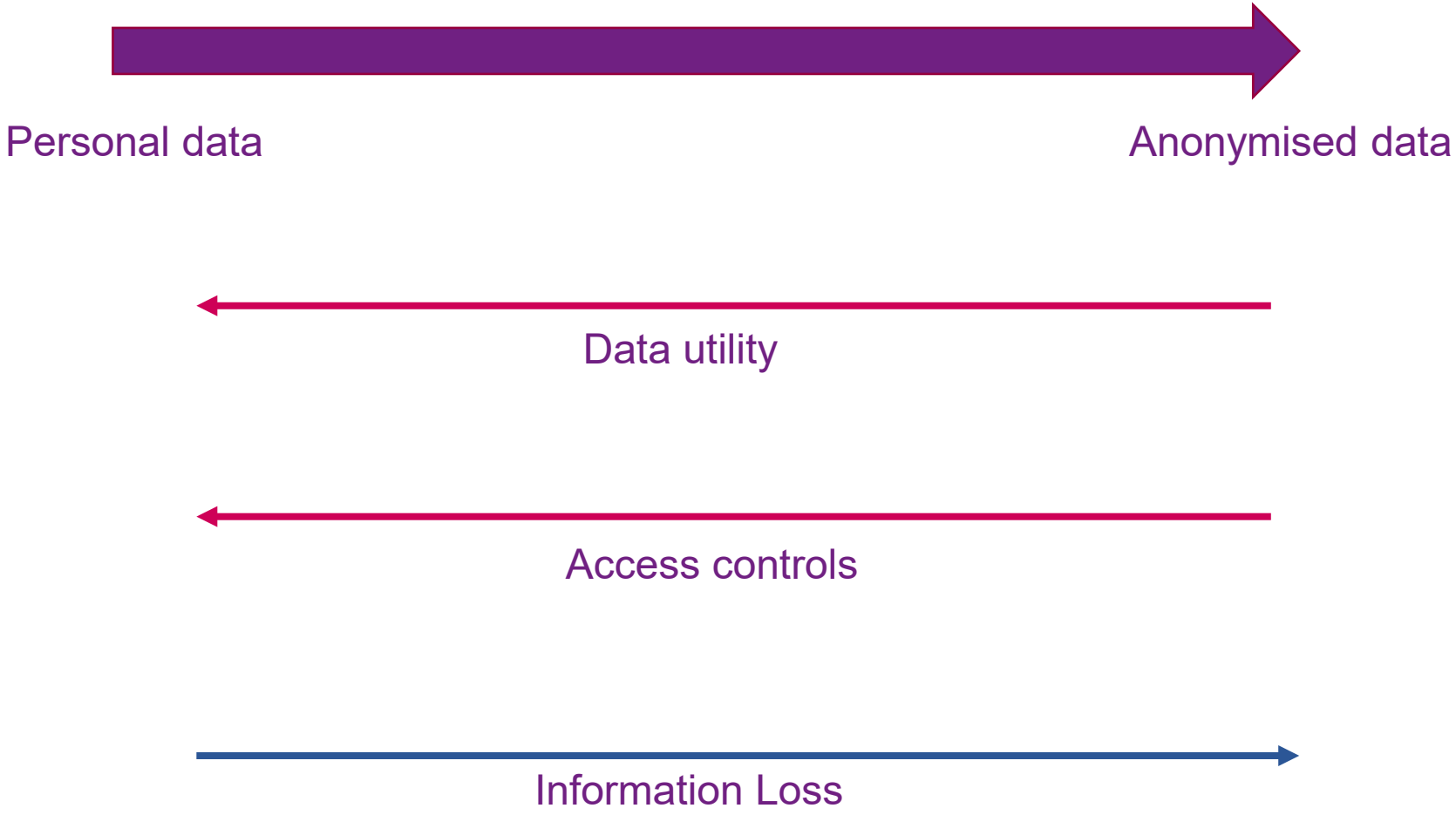
De-identification and anonymisation

De-identification of data
… refers to the process of removing or masking direct identifiers. This means that any identifiers that might directly lead to the re-identification of data subjects such as their names and their addresses are either replaced by pseudonyms (fake names), codes or removed.

Note: where only de-identification is applied data remains identifiable, hence data protection legislation applies to any additional processing of the data.

Anonymisation
… refers to the process of ensuring that the risk of identifying a data subject is negligible. This usually involves more than de-identification and often requires that data be further altered or masked, generally by treating indirect identifiers.

Information, Access, and Utility



Effective anonymisation

The ICO introduced a key concept in ensuring the utility of data, and this is effective anonymisation. They provide guidance on determining whether information qualifies as personal data or anonymous information, affecting the applicability of data protection laws. The assessment involves careful consideration of the risk of identifiability and the factors influencing it.

Data can be classified into several categories based on how identifiable individuals are.

- **Directly identifiable data**: personal data.
- **Indirectly identifiable data**: personal data, if the risk of identifiability is not sufficiently low.
- **Data unlikely to be identifiable**: risk of identifying an individual is considered sufficiently low.
- **Data that is impossible to identify**: genuinely anonymous information.

Source: ICO (2021) *Chapter 2: How do we ensure anonymization is effective?*

Anonymisation in practice

Personal data

- Anna Thomson (she/her), 45, went to her chemotherapy treatment on 5 April 2020, at Bakersfield Hospital

- Charlie (she/her), 45, went to her chemotherapy treatment in April 2020, at Bakersfield Hospital

- Charlie (she/her), 45, went to her chemotherapy treatment in April 2020, at a hospital in Oxfordshire

Anonymised

- Charlie, 45-54, went to her chemotherapy treatment in 2020, at a hospital in England

Indirect identifiers - context matters

When planning anonymisation the selection of indirect identifiers must be considered in relation to the subject of your data and the sample characteristics.

One needs to balance data utility against potential information loss and consider,

how access controls can be used.

Indirect identifiers – age

Age / date of birth information • Possible formats: • Full date of birth • Day of birth • Month of birth • Year of birth • Age (single year) • Age (in months) • Age (banded)	Single year of age is often considered one of the most important demographic information present in data. Always consider that anonymisation techniques can render some numeric analyses impossible, for example calculating the mean. Anonymisation of other variables in the file may make more detailed age variables such as year and month of birth less disclosive.

Indirect identifiers – education

Education information • Highest level of education • Field of study • Duration of study • Type of institution	Like any other indirect identifiers care must be taken when detailed information is provided. This detailed information could be unique or only applicable to a small number of people in the population. This type of information, especially when used in combination with other information that has been collected and is made available, could lead to re-identification of a participant. Where possible education information should be **categorised using a coding frame**. The usefulness of the information for secondary analysis should be considered with multiple versions of appropriate access made available where needed. For example, **educational attainment or qualification level** might be more useful than specific institutional information. The ONS Census 2021 'Highest level of qualification' codes should be considered.

Indirect identifiers – employment

Employment information • Occupation classification • Job role/title • Type of work • Employment sector • Industry sector • Employment type	Employment information is very similar to education information and the same considerations must be taken – where very detailed information is made available unique information or information only applicable to a small number of people in the population. Unique job titles may result very easily in identification, either in isolation or when combined with other information. It is recommended that employment **information is categorised and coded.** For example, the SOC2010, SOC2020 or NS-SEC coding frames can be used. Depending on the data and what other information is available responses at more detailed levels of these schemas may also need aggregation.

Indirect identifiers – income

Income and other financial information • Income in pounds • Income (range) • Savings in pounds • Savings (range) • Debt in pounds • Debt (range)	Income and other financial information can lead to re-identification especially if **unique outlying values** are presented. For example, isolated cases of a **very high or low** income or other financial sums – for example, a large lottery win recorded as unearned income - may present an increased disclosure risk when analysed alongside other information available in the data plus other publicly-available information (some lottery wins are well-publicised). Financial information needs to be carefully checked and ensure the level of detail provided does not compromise the identify of participants. Similar to the age, some statistical analyses where individual numbers are required, will not be possible, so data usability should always be considered.

Indirect identifiers – geographical information

Geographic information	Geographical or spatial variables present in the data should be considered carefully.
• Countries • Regions • County • Town • Local authority • Health area • Postcode sector • Postcode district • Full postcode • Grid references • Latitude and longitude	Detailed, low level geographic variables will exponentially increase the ability to identity participants. Always consider the **characteristics of the study**, the **sample size** and the **individual data** you are anonymising. For example, in a politics-related study, low-level political geographies are of key importance for secondary analysis. Further anonymisation can be applied to other indirect identifiers. It is advisable that for any geographical information **more detailed than regions** for **stringent access control** to be considered.

Indirect identifiers – dates of life events

Dates of life events information • Date of marriage • Date of adoption • Date of divorce • Date of death • Date of treatment • Date of court appearance	Similar to date of birth information, exact dates of life events may increase the risk of potential identification. Paying careful attention to life events is especially important as some of this information is also available in open sources that can be used to identify a participant. **Reducing exact dates to month and year** might remove enough detail when considering other information available. In other cases, reducing the exact dates to **year only** might be needed. As always it is important to assess how important the information is for secondary research. Changes will decrease **data usability** so increased **access conditions** should be considered for more **granular information.**

Indirect identifiers – ethnicity, national identity and religion

Ethnicity, national identity and religion information • Ethnic group • National identify • Religious affiliation	Detailed information about ethnicity, national identity or religious affiliation can be potentially problematic when again used in combination with other variables. Detailed responses often include unique cases. **Using a standard coding frame** is helpful and the ONS provides some guidance on this complex area. Measuring equality: A guide for the collection and classification of ethnic group, national identity and religion data in the UK Similar to other indirect identifiers the context of data sharing and usage for secondary research must be carefully considered. **Stricter access control** can be utilised to provide **more detailed information**.

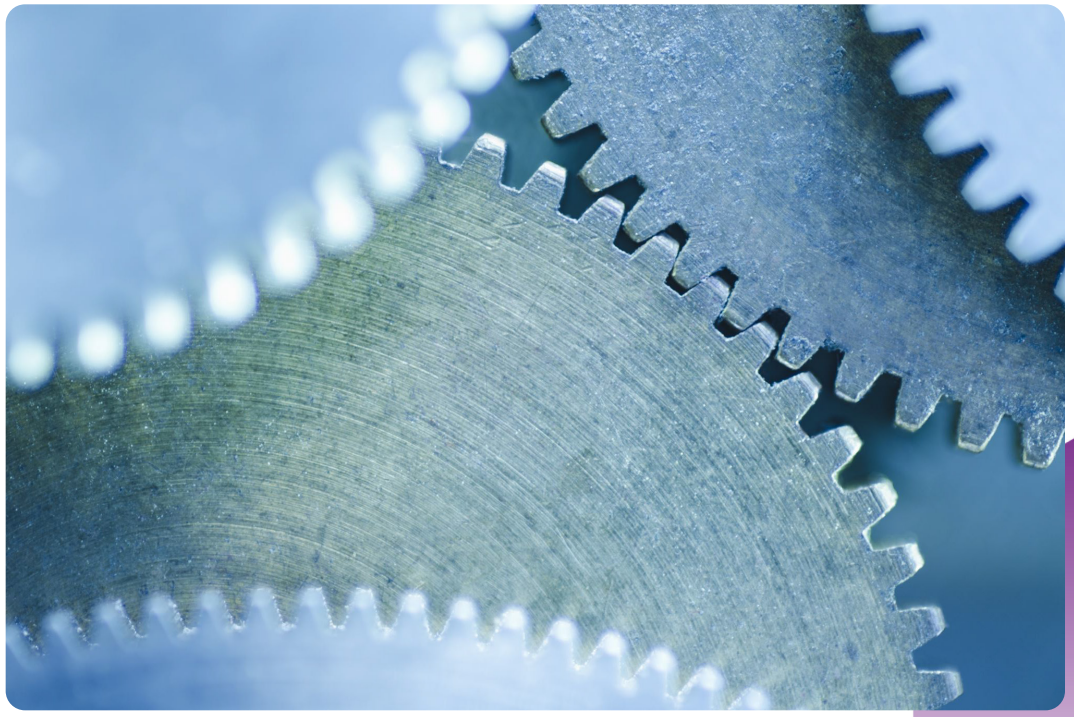
Indirect identifiers – sensitive information

Sensitive information • racial or ethnic origin • political opinions • religious or philosophical beliefs • trade union membership • genetic data • biometric data • information concerning health • information concerning a person's sex life • sexual orientation	Data should always be checked for sensitive information (special category data). Some information may include details that could be potentially harmful, if a respondent is identified, such as an unusual or sensitive health condition or status, or details of illegal behaviour. Some sensitive data will be Special Category data under GDPR and these variables should be considered alongside all other indirect identifiers in the dataset in case a disclosure risk is identified. They may **need to be edited**, but, if doing so would compromise the usability of the data, **stringent access restriction** should be considered.

Combination of information

When anonymising data the combination of information, the data utility and the access to the information must always be considered.

We are not simply editing one piece of information in isolation but considering the entire context.



Participant profile exercise

Imagine this is a row in your survey data or information from a transcript which you must share for future reuse.

You must ensure that analysis of career progression in the UK at a region level with key considerations for years of employment, highest level of education, and occupation is possible.

Which identifiers might you change and why?

Participant A

Age: 21
Date of birth: March 12, 2002
Ethnic group: Mixed/multiple ethnic group: white and Asian
National identity: British
Highest level of education: PhD in Computer Science and Informatics
Job title and job information: Senior Cybersecurity Specialist at TechSecure Ltd
Years of employment: 1 year and 4 months
Salary: £74,271
Postcode: CO1 10RP

Participant profile anonymised

Participant A

Age bands: 20-25

Ethnic group: Mixed/multiple ethnic group

National identity: British
Highest level of education: PhD

Occupation (SOC2020): Information Technology Professionals – code 213

Years of Employment: 1 year and 4 months
Salary Range: £70,000 - £80,000
Region: East of England

Data anonymisation steps

3 Steps

1. Find and assess identifiers
2. Implement anonymisation techniques
3. Review the data and re-assess any remaining disclosure risk

UK Data Service

Anonymisation techniques for numerical data

Anonymisation techniques for numerical data

Before making changes to data, remember that any changes should be made alongside data-sharing plans.

Future research needs and access levels will dictate the level of anonymisation needed and **avoid under- or over-anonymisation**.

Consider how the data are going to be shared and the legal and ethical controls in place. This will influence the appropriate level of detail required to mitigate disclosure risk present in the data.

The most common anonymisation methods for quantitative data are:

recoding, banding, top/bottom coding, generalisation.

Recoding or categorisation

We use this method to reduce the number of distinct categories or values of a variable, thus reducing disclosure risk.

Example:

Ethnicity
Black (Caribbean)
White (Irish)
Black (African)
Asian (Pakistani)
White (Scottish)
Black
White

Ethnicity
Black
White
Black
Asian
White
Black
White

Banding or binning

Banding, also known as binning or grouping, is an effective anonymisation technique commonly applied to continuous variables like age or income. It involves categorising these variables into broader ranges or bands.

Example:

Income (£)
88,599
21,478
9,996
51,299
120,987
197,000
27,998

Income (£)
75,000 – 99,000
10,000 – 24,999
Less than 9,999
50,000 – 74,999
100,000 or more
100,000 or more
25,000 – 49,000

Top/bottom coding

By top/bottom coding, the goal is to reduce disclosure risk added by small counts in the tails of a distribution.

Example:

Age
27
118
89
56
48
31
57

Age
27
80+
80+
56
48
31
57

Generalisation

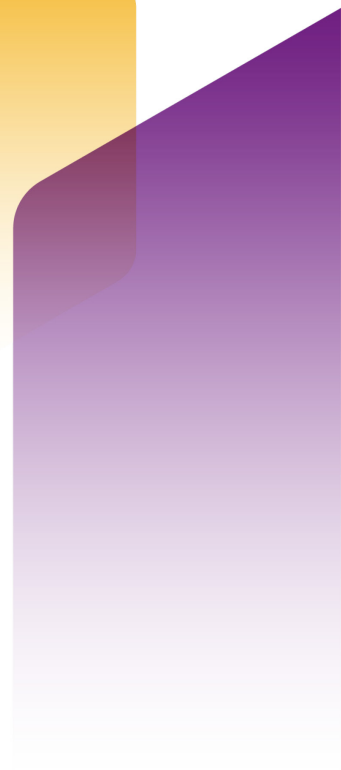
Survey data might contain free-text responses that increase the disclosure risk. To reduce the risk of disclosure you can generalise the meaning of a detailed text variable by replacing the potential disclosive information. This is a common anonymisation technique for transcript data which we will discuss shortly.

Example:
Original Detailed Response: "I am responsible for leading the cybersecurity initiatives, specifically focusing on blockchain technology security and data encryption techniques.“

Generalised Text or Coded Response: "Involved in IT security and data protection."

UK Data Service

Considerations and anonymisation techniques for qualitative data

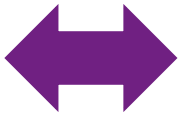


Considerations for anonymising qualitative data

Qualitative data pose particular challenges regarding anonymisation.

Considerations need to be given to the level of anonymity required to meet the needs agreed upon during the informed consent process.

Removing/redacting too much information reduces the value and integrity of the data.



Removing/redacting too little information might lead to potential identification of respondents.

Anonymising transcript data – techniques

Direct identifiers → use pseudonyms

- Direct identifiers should be disguised by allocating pseudonyms, or fake names/details, which should be consistent throughout the project, including in publications, while avoiding identifiers that could still hint at participants. Update all files, including transcripts, summaries, and metadata.

Indirect identifiers → categorise and generalise

- When anonymising transcript data, it is crucial to handle indirect identifiers (pieces of information that, while not directly revealing an individual's identity, can potentially be combined with other data to lead to the identification of a person) with care. Categorisation and generalisation are useful tools.

Anonymising transcript data – techniques (continued)

Indirect identifiers – categorise and generalise (continued)

- **Categorisation** serves as an effective strategy to preserve the utility of data while better-preserving privacy. By grouping similar data into categories, researchers can maintain essential analytical value without exposing specific details. For instance, ages within the text can be classified into ranges such as [20-25] years, [26-30] years, etc., rather than making available exact ages. Similarly, occupations can be categorised based on an international standard classification, allowing for a broader but still meaningful analysis.

- **Generalisation** is another technique where specific details are replaced with broader concepts. This approach can apply to various data types, including geographical locations, where specific places might be generalised to regions or countries. For example, generalising from "living in a small village near the town of Blackburn in Lancashire." to "living in a rural area in the North West of England".

Anonymising transcript data – best practice

Anonymisation plans

- Anonymisation plans outline the strategy in place for protecting participants' identities. They should outline what characteristics are automatically de-identified and what other potentially disclosive information to look out for.

Use find-and-replace

- Use find-and-replace tools to ensure all text is updated, accounting for variations in capitalisation, punctuation, and pluralisation. Utilize advanced search options like wildcards (? for single characters, * for multiple) to catch all permutations.

Identify changes with brackets

- Identify replacements in text clearly. A common way to highlight this is with [square brackets] or XML tags (e.g. <anonsec>anonymised text<anonsec>). Transcription guidelines should document which annotations denote changes in the text.

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

Case Study:
Pioneers of Social Research (SN 6226)

- Contains 43 life story interviews with well-known social researchers.
- Interview topics covered events and details that could be easily cross-referenced for disclosure (e.g. publications or specific research projects).
- Data sharing was facilitated with clear consent to archive personal data, and discussions with participants were highly informed (not least because the participants were researchers themselves).
- We still anonymised when we felt it was ethically necessary.

Considerations for image and audio data

- Person's image and voice is also their personal data, regardless of whether they are in a public or private space
  - This includes ethnographic work.
- The best strategy for processing, storing and sharing image data is to do so with consent
- Where anonymisation is necessary:
  - Where possible, advise participants to avoid wearing/saying anything which directly identifies them
  - Be aware of identifying context (e.g. tattoos, jewellery, identifiable background, etc.)
  - Use software which works locally on your computer (e.g. not Youtube Studio)
  - Where possible, do user testing to check that any blurring/modifications cannot be undone through re-sharpening.

Intruder testing

*Intruder testing involves using individuals described as 'friendly intruders' to try and see if they are able to re-identify anyone in the dataset. These intruders should have some background knowledge of the data similar to that of a typical user. However, they do not need to be specialist hackers with the capability of employing advanced data exploration techniques. (*ONS, 2024)

- Not intended as a replacement for theoretical disclosure risk metrics, but as a tool to be used alongside more traditional methods.

(Office for National Statistics 'Guidance on intruder testing': www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/guidanceonintrudertesting#why-and-when-is-it-necessary)

Further Resources

- ICO - Chapter 2: How do we ensure anonymisation is effective? ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf
- ICO - Anonymisation: managing data protection risk code of practice ico.org.uk/media/1061/anonymisation-code.pdf
- ICO - Key data protection terms you need to know: Processing ico.org.uk/for-organisations/advice-for-small-organisations/key-data-protection-terms-you-need-to-know/#processing
- ONS Policy for social survey microdata www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata
- CESSDA Data Management Expert Guide dmeg.cessda.eu/Data-Management-Expert-Guide/5.-Protect/Anonymisation

Useful semi-automated tools

- 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
- QuailAnon – open-source tool developed by QualiService
- Text anonymisation helper tool – Word Macros tool developed by the UK Data Service
- Textwash - open-source tool uses Python to identify and replace direct identifiers
- FAMTAFOS – in development; open-source desktop app that utilises AI technology to anonymise text at scale; operates on principle of Named Entity Recognition (NER), and can be set to search for names, locations, occupations, etc. They will then tag them for subsequent human editing.
- De-ID - HIPPA-compliant tool to flag potentially identifiable data; only available to organisations

Get connected

UK Data Service

Jisc mail group

@UKDataService Twitter

UK Data Service YouTube channel

Powerpoint slides will be available on our website in due course and you can catch up on the recording on our Youtube channel. Check out our Twitter for more updates.

UK Data Service

Thank you.

datasharing@ukdataservice.ac.uk