UK Data Service

ukdataservice.ac.uk

Introduction to anonymisation techniques for social sciences research data – Q&A May 2025

22 May 2025

Public

Version: 01.00

“From the definition of personal data you mentioned, and especially concerning indirect identification, I understand that the external information that would allow to re-identify the person is only information that is published. My question is: does that imply that information known by an individual because, for example, they personally know the person involved in the research, does not count as information that would make the data re-identifiable?”

The slide specifies "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." which is meant to be understood as anything that relates to an identified or identifiable person that is available in the data plus any other information that is available out there. Especially with the grow of social media platforms, where people make a lot of personal data available, there needs to be a consideration around what we as researchers make available, could that be linked to other resources such as social media, or public registers.

“How can we be confident about ethical questions over and above anonymisation?”

It can be difficult to strike the correct balance. This is where a dialogue between the repository where the data are made available and the researchers that are anonymising the data is extremely useful. Anonymisation should match the communication with the participants but also match the access control for the data. It is always advisable to think from a secondary user perspective, what is needed in the data to be able to analyse, and trying to retain as much of the information as possible while using the techniques we have discussed such as generalisation, recoding etc..

“I was wondering if IP address is always direct? because: 1) people could be using VPN 2) it only links to a machine/general geographic region?”

When a VPN is being used that can no longer be tracked to one specific individual, indeed the IP address is no longer an identifier at all, neither direct or indirect. However, unless we are certain a VPN has been used is better to take a more careful approach where the IP address is treated as a direct identifier. The IP address identifies the device which can be linked with the name and account information of just one individual. This is not applicable in many cases but can be and therefore being added to the list of direct identifiers. Of course where no account is linked to the IP address, it can serve as an indirect identifier from a geolocation perspective (and where no VPN is used).

“Is there a framework to assess global level of anonymisation of a shared dataset? You mentioned the open-safeguarded-controlled three tiered levels of access: is there some similar three tiered framework to state/declare the global anonymity of a dataset?”

Anonymisation is always about the combination of the different information/variables. When it comes to numerical data a fantastic tool that can be used to assess the "global disclosure risk" is sdcMicro https://cran.r-project.org/web/packages/sdcMicro/index.html. It does offer an UI as well and only three lines of code need to be run in R to open the UI. The tool allows you to visually inspect all the rows in the data based on the calculated risk. The three tier access licensing framework is what is used at UKDS https://ukdataservice.ac.uk/help/deposit-data/deposit-in-the-curated-data-repository/curated-data-repository-licensing-and-access-framework/ other repositories might use different terminology.

There is no global unified framework yet either when it comes to access/licensing levels or risk of disclosure, however there are efforts, driven by research councils to gain a unified framework.

“Is race/ethnicity always sensitive? For instance, certain parts of London has high ethnic communities, so wouldn't in those places it's actually the norm (i.e. not uniquely identifiable)?”

Under UK GDPR personal data revealing racial or ethnic origin is special category data, therefore considered sensitive data. It is true that ethnicity is commonly discussed and may be perceived as less or not even at all sensitive, however this does not change the legal status of the type of data therefore the additional considerations still stand. This does not mean that this type of data cannot be collected or made available for secondary reuse, it just means more care needs to be taken, including identifying and additional legal basis for processing the data as discussed, and applying anonymisation techniques and access control to the data that is being shared.

“Is data of a deceased person by definition anonymous? If yes, "date of death" should not be

considered as an indirect identifier but anonymous?”

Very true, UK GDPR applies to living individuals however the duty of confidentiality still stands and date of death can be used very easily to link with other data and therefore be able to identify the participant. Therefore extra considerations are necessary for date of death and is considered an indirect identifier.

“As I understand it this webinar has been about anonymising for sharing/making data open. But if and how is this different for data you just keep in your HEI (e.g. PhD data), do you have any comments on working with data when not sharing it?”

Data privacy laws pertain to any personal data that you are processing – whether that processing is sharing the data or simply collecting/analysing the data. Therefore, an appropriate anonymisation strategy and any other considerations (including regulating access and/or obtaining informed consent) should be planned before collecting data, so that any processing of the data is done so lawfully and ethically. If you are not planning to use a responsible repository for the long-term preservation and sharing of the data (e.g. in the example of PhD data), then you should ensure data is stored and/or destroyed in accordance with the information told to participants at the point of getting their informed consent and your institutional policies. All institutions have retention schedules.

“Linked to the ‘definition of personal data’ question – if you are interviewing someone about, say, their marriage, and you remove all info that could make them identifiable to the general public – but their spouse could still tell it is them, is that anonymous? That is just an example, but my query is about whether data (e.g. interview transcripts) that would not be identifiable to the general public but would be identifiable if their friends or family or colleague read them.”

This is partly what makes qualitative data so tricky to handle – the rich detail that makes qualitative data so useful is also what makes it difficult to assess disclosure risks. Who knows what information, and what is the likelihood that those individuals would seek and/or find this data? This is where 2 key concepts from the online workshop are particularly useful:

1. The 3-prong strategy – of informed consent, appropriate access restrictions, and anonymisation – give additional safeguards where it is difficult to assess the disclosure risk of the data. The informed consent ensures that you are treating the data in accordance to the wishes of participants. During this process, you should inform participants what will happen to the data after the project and who has access to the data at any given point. (Note: Consent is not usually the legal basis for processing the data, but can provide ethical guidance on how to treat the data.) Appropriate access restrictions should be applied at all times. While the project is on-going, participants should be made aware of who will have access to the data, including any collaborators, partners, and auditors of the data. If depositing the data in a responsible repository, there is also scope to adjust access to the data by choosing a license which will set the level of access appropriate for your data. Anonymisation of direct and indirect identifiers then also layer onto these decisions.
2. The ICO has used the term "effective anonymisation" to describe where personal data has been assessed and removed, and it is reasonably unlikely that individuals would be recognised. Moreover, in the event of a breach, damage can be mitigated. It's sometimes difficult to know for sure who knows what information (although we assume someone will know/remember life events described in interviews), so it feels safe to take the more risk averse approach in assessing disclosure risk. However, this is all about likelihood: what is the likelihood those individuals would be able to access this specific dataset and identify that person?

"Can you give further details on the occupation coding tool, which I assume translates free-text occupations into 4-digit Standard Occupational Classifications (2020). Is this CASCOT, or a different tool?"

This coding tool is CASCOT – available online here: https://cascotweb.warwick.ac.uk/#/classification/soc2020. There may also be a downloadable version for your computer available. There is also guidance on the categories that are used, available through ONS.

"I plan to interview NGOs and the local community about an unethical company. There is only one big company in that region. Considering the risks for the employees and other aspects (e.g. risk on stock price) of the company, should I anonymise the location by mentioning the country only? Do you have any experience with this?"

This is a tricky one to assess – based on the question posed, I assume you want to protect the company name here (perhaps because it becomes disclosive to individuals working there?). We have had collections where the location matters and is recommended not to be anonymised. For example, we have a collection on trade union activity for ship-builders. Another example is a study on the foot and mouth disease and its spread through farms in England. In these cases, location could not be easily anonymised (or there was little point in anonymising the location), so other details were anonymised instead to ensure individuals could not be easily re-identified. I would ensure that you have a good discussion with participants about what you plan to do with the data and what kind of outputs you plan on producing and get their views on the best way to protect identities. The other option, of course, is to aggregate the location to higher level. If region in still too narrow, then you can perhaps generalise further (e.g. North of England or South of England).

If you want to list the company name in this research, it's worth pointing out that laws on personal data applies to individuals; the Statistics and Registration Services Act does apply to personal information, but it does not apply to individuals researchers handling sensitive data not designated for official statistics.

"I work in qual data. Often we have transcripts where we remove identifiable info, but the individual has a unique ID number and in a separate document the ID number links them to direct identifiers. As I understand it, until that second document is destroyed, the data are not anonymous. Is that correct? Is there anything else I need to know when working in this way?"

Having a separate listing linking the participant IDs to their real information would usually be considered pseudonymised data (note: we did not discuss pseudonymised data in the online workshop because there are some variations in how this term is defined, and the focus of anonymisation is to produce data where UK GDPR no longer applies; i.e. personal data is no longer present). The ICO has published some guidance about pseudonymisation as a technique to help maintain security and reduce risk, but notes that "you **should** still consider whether you can meet your objectives by using anonymous information". That data with the personal data still present would need additional safeguards – beyond what you have done for the rest of the data – which is called the "security principle". You are taking on additional risks to yourself and your organisation if there is a breach of security, and the ICO would examine whether sufficient security was in place for the storage of that personal data.

To further note here is there are situations where pseudonymise data needs to exist, for example in the case of longitudinal studies, quantitative or qualitative. When you share the anonymised data and you retain the key that is not made available to those having the anonymised data, that data is anonymised in their hands.

"How can interview transcripts ever be effectively anonymised because of pattern of speech, certain sentence structure, phrases and ways of describing things, can be analysed and may be linked to individuals?"

The nature of qualitative data does make this type of data much harder to assess for disclosure. Sometimes, there may be certain features of speech which – if transcribed phonetically and/or verbatim – could increase the risk of a disclosure. There are options in transcriptions to hide this. While many rely on verbatim transcriptions, you can also choose to use an edited transcript style, where some of these features may be masked. Additionally, we have had some collections use summarisation tools to summarise the transcript and release interview summaries at a more open (or safeguarded) level of access, while the transcripts are left at a more restrictive "permission only" access. Where these speech patterns are not essential to the original project, we would encourage researchers to consider using an edited transcription style.

The three prong strategy as well is a key consideration as this needs to be explained to participants, data is anonymised and then access control is put in place to ensure that re-identification is not possible.

"If a participant mentions a public figure and this is relevant to the context should this be replaced (e.g. pop group, politician)?"

This may depend a bit on the context of your research project, so it's worth getting in touch about the specifics where you are worried about a potential disclosure. Very generally speaking, for most data collections, I would normally say you probably don't need to anonymise names of individuals are understood to live/work in the "public eye". Having said that, it depends on the context: having an opinion about a politician's public speech or noting you like Taylor Swift's music, for example, may be not constitute a disclosure. Conversely, discussing a personal relationship with a well-known actor, however, may constitute a disclosure. Context matters here, so if you are in doubt, please get in touch for more specific advice. It is also important to bear in mind whether defamation would be applicable in this context, while separate from disclosure, it's still important to consider.

"How do you provide anonymisation to experts when you need to provide context for their expertise and credibility to research, i.e. providing evidence of their expertise?"

In cases where niche areas of expertise are discussed, it is very difficult to anonymise this information. The case study used in our online workshop, the Pioneers of Social Research, is an example of this: 54 life history interviews were conducted with eminent scholars. In this case, explicit permission to publish these interviews openly was sought before conducting the interviews. We still assessed these for areas where ethically we need felt a need to lightly edit transcripts (e.g. where court cases, unfavourable opinions of others, and/or medical issues of those not involved in the study were discussed), but did not remove personal data of the interviewee. If you do not have explicit permission from participants to use their names, or they have requested that you do not, you'll need to carefully consider your anonymisation strategy and access restrictions on the data. Where consent was not originally sought for this, consider whether you can get retrospective consent to publish data. We did this as part of the digitisation for another data collection, the British Oral Archive of Political and Administrative History, where individuals in Winston Churchill's administration were interviewed. Since permission to publish this data online was not sought at the point of data collection (this simply was not an option at the time this study was conducted), we contacted individuals (if living) and families of the individuals (if deceased) to ask for explicit consent for release of this information. While consent is considered an ethical element here, rather than the legal basis for processing the information, it can ensure that your treatment of the data respects the wishes of those involved.