ADR UK feedback to ICO on the draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, chapters 1 – 3

About ADR UK

ADR UK (Administrative Data Research UK) is a partnership of government and academic groups across all four UK nations (ADR England, ADR Northern Ireland, ADR Scotland and ADR Wales) and the Office for National Statistics (ONS). The partnership is coordinated by a UK-wide Strategic Hub within the Economic & Social Research Council (ESRC), which is part of UK Research and Innovation (UKRI).

ADR UK is creating linked research datasets from administrative sources, making these available to researchers through our network of trusted research environments (TREs). Administrative data is created when people interact with public services that keep records to carry out their day-to-day work. Although not originally created for research, this data has great potential to provide insights to help policymakers and others make better informed decisions. ONS is our major data infrastructure partner, with ADR UK directly funding the development and expansion of the ONS Secure Research Service (SRS). We also fund TREs in Northern Ireland, Scotland and Wales, respectively run by:

- NISRA
- the Scottish National Safe Haven
- Swansea University (the SAIL [Secure Anonymised Information Linkage] Databank).

ADR UK started in 2018 as a pilot programme, where we tested different ways of working with data owners and researchers. As a result of the success of this pilot, in 2021 we had confirmation of £105 million of long-term funding up to 2026. This decision was approved by ESRC, UKRI, the Department for Business, Energy & Industrial Strategy (BEIS) and Her Majesty's Treasury (HMT).

Part of the success of the pilot was our collaboration with other UKRI investments that are also supporting opening up data for research, such as Health Data Research UK **(HDR UK)** and the National Core Studies programme. Through these collaborations, ADR UK has been able to open up access not just to administrative data for research, but also health data – for example, via the ONS SRS, the SAIL Databank and Research Data Scotland. This is because ultimately, the design of a secure, robust TRE for administrative research works equally well for health data, which then opens up the potential for linking health and administrative data, as the ADR UK partners in Wales and Scotland have been doing for many years now.

As we move out of the critical phase into the recovery phase of the pandemic, ADR UK will be playing an increasingly pivotal role, as the need for insights based not just on health data, but on education, benefits, income, homelessness, crime and justice data grows.

The ADR UK model for working with data owners and researchers is all about bringing government and academic groups together into collaborative partnerships. The aim is to deliver policy-relevant research that reinforces the feedback loop between those who have collaborated with us to open up access to data and the researchers commissioned to analyse it. Our model speeds up access to data by setting up partnerships between government data owners and external researchers to create the linked datasets, which are then hosted within our TRE network. Any external researcher can apply to access them for research that is in the public interest.

Since ADR UK began and the Digital Economy Act was fully enacted, both in 2018, the combination of having funded infrastructure and legislation in place means that research based on administrative data is now making a significant impact across a wide range of research sectors, increasing the bodies of knowledge in these areas, and informing policy practice.

ADR UK's response to this consultation is drawn from the perspective of using administrative data for research, and the issues around anonymisation, pseudonymisation and de-identification related to this work. However, we recognise that different types of research require different levels of anonymisation.

General thoughts

ADR UK welcomes the ICO's decision to create guidance on anonymisation, pseudonymisation and privacy enhancing technologies. The draft guidance is technically sound, accurate, and a useful addition to the field. However, we question whether the guidance in its current form is sufficiently accessible and easy to apply, especially for practitioners without much experience in anonymisation. This issue might be resolved with additional guidance from the ICO, for example, through training and infographic materials.

We recommend that the ICO takes care to fully and accurately distinguish between 'anonymisation', 'pseudonymisation' and 'de-identification' throughout the guidance. This distinction is crucial because there is a risk involved in using 'de-identification' to refer to any kind of anonymisation. De-identification – or removing direct identifiers – of data is a relatively weak form of anonymisation, and not suitable for every scenario. If data owners are led to believe that these concepts are interchangeable, then they may become reluctant to share their data – as de-identification alone may not provide a sufficient level of security. We have suggested points of clarity on this distinction throughout our response.

As a whole, the guidance seems geared towards quantitative data. Qualitative data is also valuable for research, and suitable guidance is also needed.

Overall, we are pleased to see that the guidance advocates for anonymisation as an effective means of protecting data. This will help to avoid the polarisation of data access as either 'open' or 'secure', with nothing between these two states.

Chapter 1 – Introduction to anonymisation

This is a helpful framing of the broad issues that need to be considered when anonymising personal data. It is also good to see the ICO recognising the wider benefits of sharing data to support research. Specific feedback on some points where further clarification would be helpful are presented below.

From the perspective of ADR UK, we feel this sentence on page 10 could be presented in a more nuanced way:

"Clearly, 100% or 'absolute' anonymisation is the most desirable position."

As you acknowledge, true anonymisation of data – so, not just the removal of personal identifiers, but the consideration of whether individuals can be indirectly identifiable when data are linked – can be technically challenging and complex. It also may not always be appropriate or possible, as it depends on knowing what other information could feasibly be accessed that would allow re-identification.

While there may be incentives for organisations to only allow access to data in an anonymised form (for example, the data will no longer fall within the scope of the General Data Protection Regulation [GDPR]), this may devalue the data to such an extent that it is no longer useful for research purposes. The context of the data disclosure must also be taken into account – and whether data privacy or informational privacy is the more relevant consideration. So, before anonymisation is contemplated, we would strongly argue that the utility of the resulting data, if it is to be used for research purposes, needs to be considered. Also, organisations need to have sufficient skills, resources, technology, and processes in place to undertake this work effectively.

Data de-identification, referring to the likelihood of data subjects being re-identified directly from that data, is a much more useful concept when considering the research use of data. If de-identified information is accessed via a trusted research environment that is accredited under a suitable process, for example under the Digital Economy Act (DEA), and as such meets all the required security standards, then the de-identified data can be considered functionally (or effectively) anonymised, through a combination of the actions taken to create the de-identified dataset, and the environment in which the dataset is accessed. As an example, the design of a DEA-accredited trusted research environment means researchers are not allowed to bring in other sources of data, then attempt to link these with the de-identified data being accessed within the trusted research environment, to re-identify data subjects. Therefore, the residual risk related to the creation of a de-identified dataset rather than an anonymised dataset is mitigated and the data retains its value as a research asset.

For these reasons, we would argue that 'absolute' anonymisation is not always the most desirable position. In the context of ensuring data retains its use for research purposes after being de-identified, this statement could also be quite unhelpful, as worded. We therefore suggest the ICO amend this statement.

On page 15, the draft guidance states:

"… for the purposes of the re-identification offence, the DPA 2018 refers to 'de-identified' personal data as personal data that has undergone pseudonymisation as defined in the UK GDPR rather than (for example) anonymous information."

Then:

"Pseudonymisation means that individuals are not identifiable from the dataset itself, but can be identified by referring to other information held separately. Pseudonymous data is therefore still personal data and data protection law applies."

However, as it is entirely possible to create de-identified data that cannot be identified, we ask if the guidance could be updated to clarify that pseudonymised data and de-identified data are not always the same. This is important, because as the draft guidance states:

"…Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person…"

Also:

"… personal data which has undergone pseudonymisation remains in scope of the law…"

If de-identified data is accessed via a TRE that is accredited under a suitable process (for example under the DEA) and, as such, meets all the required security standards, then the de-identified data can (and should) be considered functionally anonymised, through a combination of the actions taken to create the de-identified dataset, and the environment in which the dataset is accessed. As such, this data is created specifically so that it cannot be re-identified. It would be helpful if the ICO could consider including an example that presents this as a way of lawfully making data that was originally personal data accessible for research purposes – particularly as functional anonymisation is a concept that sits well with both data owners and the public.

On page 17, there is a reference to pseudonymised data being useful in the context of processing personal data for:

"scientific, historical, and statistical purposes."

It would be helpful if this referred to **research** purposes. This would ensure easier read-across from the ICO's guidance to that relating to the use of the DEA as a legal gateway for supporting data sharing and access, as the DEA refers to research and statistical purposes.

Editorial comment on Chapter 1:

On page 10, this sentence is missing an 'a':

*"However, this residual risk does not mean that **a** particular technique is ineffective."*

Chapter 2 – How do we ensure anonymisation is effective?

This is a helpful framing of the broad issues that need to be considered to ensure anonymisation is effective. Specific feedback on some points is presented below.

On page 7, the draft guidance states:

"To determine the likelihood of identifiability through inference, you need to consider the possibility of deducing the identity of individuals from:

… other information that you either possess or may reasonably be expected to obtain. For example, this could include publicly available additional information, such as census data."

We suggest that referring to Census data in the context of publicly available data that could be used to re-identify data is not helpful since Census records in the UK are not released to the public until 100 years after a Census took place. As this sentence is currently drafted, the reference to Census data could easily be misinterpreted to mean more recent Census data, which is only accessible to accredited researchers working within DEA-accredited TREs on accredited projects, after it has been de-identified. The risks related to the re-identification of datasets in the context of publicly available Census data is very different if the only available Census data to link to are over 100 years old. As an alternative to referring to Census data, perhaps the ICO could refer to electoral roll data?

On page 9, the figure refers to data being 'impossible' to identify. This is a very strong term; is it helpful in this context, if it motivates people to attempt to re-identify data classed as 'impossible' to identify as a technical challenge, under the cover provided to security and technology researchers acting in the public interest, then publish exactly how they did it? Surely the impossibility of the re-identification is only defined by the skills and means available to those motivated to attempt it? Perhaps 'extremely unlikely' would be a more helpful phrase?

The phrasing around 'impossible to identify' also suggests an unhelpful approach of '[privacy protectionism]' – where the focus of data protection falls too much on controls rather than safeguards, to the detriment of the data.

Linked to this point, as you reference on page 11, Recital 26 of the UK GDPR states:

"To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used… to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely… account should be taken of all objective factors… taking into consideration the available technology at the time of the processing and technological developments."

Referring to the impossibility of data identification as a desirable (or achievable) outcome appears to put a very high level of responsibility on the data owner to understand the available technology at the time of the processing and technological developments. This may not be helpful in the context of the purpose of this guidance. Also, as the draft guidance acknowledges further on:

"It is not always possible to reduce identifiability risk to a level of zero, and data protection law does not require you to do so."

On page 11, the draft guidance introduces the term 'effectively anonymised' to describe a state where the chances of re-identification have been sufficiently reduced to allow the data to be treated as non-personal data, in terms of the law. It would be helpful to include a note that this state is described by many organisations (and in the published literature) as 'functional anonymisation' – as discussed in our response to Chapter 1 – and that the two terms are synonymous.

On page 11 you invite readers to:

"...consider the means reasonably likely to be used at the earliest stage of your anonymisation process, particularly when deciding the "release model" (i.e. public release, release to defined groups etc)."

It would be helpful to include an example of a DEA-accredited TRE as a release model, as the security that is built into such a release model mitigates the risks around re-identification of data. As such, it is a constructive example, underpinned by law, to present to data owners considering how to safely and securely make their data accessible for research purposes. The context of the security of the environment through which data access is given could also be added to the flow diagram on page 27.

Editorial comment on chapter 2:

On page 4, there is an extra full stop at the end of this sentence:

"However, as detailed above, the definition also specifies other factors such that can mean an individual is identifiable.."

Chapter 3 – pseudonymisation

This is a helpful framing of the broad issues that need to be considered to ensure pseudonymisation is effective. Specific feedback on some points is presented below.

On page 10, there is reference to:

"… undertake further processing for archiving, scientific or historical research, and statistical purposes, which are automatically considered to be compatible purposes"

The specific reference to research is helpful, particularly in the context of feedback given on Chapter 1 (page 17), where there is a reference to pseudonymised data being useful in the context of processing personal data for 'scientific, historical, and statistical purposes' (but not research purposes). It would be helpful if both referred to research purposes. This would ensure easier read-across from the ICO's guidance to that relating to the use of the DEA as a legal gateway for supporting data sharing and access, as the DEA refers to research and statistical purposes.