Anonymisation and pseudonymisation of research data

In this document we explore how anonymisation and pseudonymisation applies to the sharing and publication of research data.

# Introduction

It is fair to say that anonymisation is a complex topic. There are three key points to bear in mind as you work your way through this post, they are:

1. Anonymisation is a**process** which essentially **renders personal data non-personal**.
2. To render personal data non-personal, you will need **to assess and manage re-identification risk**.
3. To assess and manage re-identification risk, you will need to consider the **relationship between data and the environment in which they exist**.

In this document we look at: why anonymisation should be considered at the study planning stage, key terminology, perspectives on anonymisation and, introduce a framework for doing anonymisation.

# The importance of anonymisation in research practice

Anonymisation is a process that allows research data to be shared and or published safely and responsibly.

You may think that setting out your approach to data sharing and publication and in particular anonymisation (to support these activities) is, at the study planning and set-up phase, a bit early. It isn’t. This is because all research activities need to be transparent in order to be considered fair, and in turn not to impact on what it is you plan to do during your study and beyond. Let us consider this further; anonymisation is a form of data processing (because you are rendering personal data non-personal) and as such your plans around this activity need to be transparent. How you do this in practice, is to include details (in your public facing documents) on what data you plan to share and publish and on how you will manage data confidentiality.

The landscape in which you as researchers will publish and share data is complex. There is an increasing amount of large and linked datasets being created about people. Couple with this, technological advances enable data to be exploited in new and previously unimagined ways. When you also consider that anonymisation is itself a complex process, it is not surprising that there have been a number of high-profile cases of re-identification of confidential datasets. Cases of re-identification of confidential datasets often involved data that was considered to be anonymised being ‘de-anonymised’, and confidential information being revealed. We will look at two examples in the next section.

## How can Re-identification happen?

It is worth noting that the re-identification problem (i.e. the re-identification of persons within a confidential dataset) is not new.

One of the most well-known cases of re-identification was of a US Governor by Latanya Sweeny in 1996.In the mid-nineties The Massachusetts Group Insurance Commission (GIC) released for research purposes a (supposedly) anonymised hospital insurance dataset of all hospital visits by state employees. The GIC removed obvious identifiers such as name, address and social security number. At the time of its publication, the Governor of Massachusetts came out in support of the published dataset, reassuring the public that patient privacy was protected.

Latanya Sweeny, then a Computer Science graduate student, set out to demonstrate that re-identification was possible in this (supposedly) anonymised hospital insurance dataset. She was not primarily interested in the information revealed (of course the revealing of information might well be a motivation of a would-be data intruder). Sweeny was able to re-identify the Governor of Massachusetts in the dataset; she knew the area he lived in and she had publicly available information i.e. a voter registration file (she purchased for $20) and media reports of the Governor being hospitalised after a collapse at a rally. Given the auxiliary information she had — she was able to match on date of birth, gender and zip code. Sweeny sent the Governor’s health record from the dataset to his office, not surprisingly this led to changes in the privacy law in the US.

Although this case is over twenty years old, it is still pertinent today as the core issue remains the same.

**When determining whether data are anonymised or not, you need to take account of:**

* **the data to be released or shared,**
* **the (release /share) environment, including what other data might exist in the environment,**
* **who is able to access that data and what – if any – governance and infrastructure controls and restriction are in placed to manage data access and reuse.**

Demonstration attacks, such as the one carried out by Sweeny, have in more recent years led to some to argue that anonymisation is a failed tool. For example, Ohm (2010)¹ argued that **‘data could be useful or anonymous, but not both’**. This position has been widely contested, with critics pointing out that incidences of re-identification (from demonstration attacks) arise because anonymisation is **poorly applied**rather than becauseanonymisation is an ineffective tool in protecting data confidentiality**.**

The re-identification of the Governor of Massachusetts by Sweeny is a case in point of poorly applied anonymisation, where too little attention was given to the interaction between the data and the publication environment.

To help you start thinking about anonymisation consider the question below and the example. You may wish to revisit your answer once you have worked through this post and look at whether your decision changed and if so, why?

## Can you publicly release this dataset?

A media streaming service called ‘AB’ want to release a movie rating dataset. The dataset contains 100 million movie ratings based on the views of half million of its subscribers. In the dataset are the following variables:

1. A subscriber ID (uniquely generated by AB)
2. Movie titles and year of release
3. Date and time on which the subscriber rated the movie

**The short answer is no.** This example is based on a real-world data release and subsequent demonstration attack of re-identification.

**Let us take a look at what happened:**

In 2006, Netflix publicly released a (supposedly anonymised) version of their movie rating dataset for the purpose of improving its movie recommendation algorithm. Netflix planned to crowd source the problem — whoever came up with the best solution to improve the algorithm would win $1 million dollars.

University of Texas Computer Science researchers Arvind Narayanan and Vitaly Shmatikov, showed that the dataset was not sufficiently anonymised for public release. They were able to re-identify a sub-sample of people in the Netflix dataset by cross matching it with another publicly available dataset, namely the Internet Movie Database (IMDb), linking on movie ratings and time stamps.

Importantly IMDb also contained at least some contributor names along with their movie rating and time of posting.

# Key terminology

The terminology associated with anonymisation can be confusing. Different terms may be used i.e. ‘anonymisation’, ‘de-identification’ and ‘pseudonymisation’. In order to apply the terminology appropriately, the first thing you need to know is how personal data is defined in data protection law.

Data protection law is underpinned by a bipartite model of data: meaning data is classed as either personal data or not (personal data), there are no in-betweens.

## Anonymisation

If data is not personal data, it is classed as anonymous information. The legislation has little to say about anonymous information other than, it is:

1. **Information that does not relate to (or no longer relates to) an identified or identifiable natural person.**
2. **Out of scope of the legislation.**

To learn more about anonymous information then, we need to understand what it means for persons to be ‘identified’ and ‘identifiable’ and this requires that we examine the definition of personal data. UK GDPR defines personal data as

“any information relating to an identified or identifiable natural person …;**an identifiable natural person is one who can be identified, directly or indirectly**,” (article 4(1))

The part of the definition of interest to this discussion is in bold, that an identified person is one that can be identified from data either:

1. **directly**

or

1. **indirectly**

We shall return to this point as we describe how the terms de-identification, pseudonymisation and anonymisation are understood in the UK and Europe. Referring to the above as condition a) and condition b).

## De-identification

The term de-identification has been used variously to describe the replacement or masking of direct identifiers such as name, address and unique (common) reference numbers.

The term ‘de-identification’ applied in this way addresses no more than **condition a)**, that is, the risk of identification arising directly from data

## Pseudonymisation

When GDPR (2016) came into effect in May 2018 it introduced a new term to data protection legislation, that of, pseudonymisation which is defined as:

**‘…** 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(5)).

Pseudonymisation, as you can see from the definition, is a process not dis-similar to the description given of de-identification, in that it addresses no more than **condition a)** the risk of identification arising directly from data. Since it would only require a look-up table or details of the encryption algorithm to break the protection. Note the sentence in the definition of pseudonymisation, to paraphrase, pseudonymisation is the processing of personal data where identification is not possibly directly from the data.

**The important point to take away from this is that within the framework of GDPR, data that has undergone the process of pseudonymisation is considered personal data.**

The process of anonymisation, in contrast to pseudonymisation, should address both **conditions a) and b)**, i.e. the risk of identification arising directly and indirectly from data. The replacement or masking of direct identifiers is necessary, but rarely sufficient on its own for anonymisation.

# How best to think about anonymisation

Although anonymisation has privacy implications it is not primarily concerned with privacy but with keeping data confidential. Within the field of data confidentiality, anonymisation is closely tied with a particular type (of anonymisation) which you may be familiar with, called Statistical Disclosure Control (SDC).

SDC covers the integrated processes of disclosure risk assessment, risk management and data utility. SDC is however just one aspect of anonymisation. Anonymisation is much broader than a set of technical processes for managing risk, it should include along with it technical processes, integrated legal, governance and ethics processes.

A common error when thinking about anonymisation has been the tendency to focus on a fixed end state of the data as ‘anonymised’. This has tended to lead to:

1. The use of success terms such as ‘truly anonymised’ (which should be avoided) and,

2. Muddled thinking about what it means to produce ‘anonymised data’, resulting in:

* An almost exclusive focus on the properties of the data whereas in reality whether data are anonymised or not is a function of both the data and their environment.
* Discussions about the relationship between anonymisation and its companion concept risk, and in particular an over optimistic assumption that ‘anonymised’ means that there is zero risk of an individual being re-identified within a dataset.
* The assumption that one’s work is done once the anonymisation process is complete and the end state is produced, which in turn promotes a counterproductive mentality of ‘release-and-forget’.

**So how should you think about anonymisation? Let’s consider the first point made in the introduction:**

‘Anonymisation is a **process** (for rendering personal data non-personal)’. More precisely it’s a process of risk management that requires dynamic decision- making.

This is because in a data flow, where data moves from one environment to another, the classification of the data as either personal data or anonymous information may change. In other words, data considered anonymised in one particular environment may not be anonymised in a different environment.

The classification of data may change depending on the level of re-identification risk associated with it, this is tied not just to the data but**who is looking at the data and the environment in which the data exists.**

The question you need to ask yourself when thinking about sharing or publishing data is ‘Are these data anonymised for persons X in environment Y?’.

# The role of anonymisation in data protection

There is an ongoing discussion on the role of anonymisation in data protection about whether or not it ought to be able to provide an ‘absolute assurance’ of data confidentiality. There are two camps of thought on this, they are:

1. **The absolute approach, which is underpinned by a belief that there should be zero risk of re-identification in an (anonymised) confidential dataset.**
2. **The risk-based approach, which is underpinned by a belief that there is an inherent risk of re-identification in all useful data.**

Now you may think that zero risk of re-identification ought to be the data protection standard that is met. However, anonymisation is not just about data protection. It is a process inseparable from its purpose to enable the sharing and dissemination of useful data.

Thus data utility (how useful the data is for the user) is another important consideration here, because there is little point in sharing data that has little or no utility for its intended audience. Low utility data may still contain a risk of identification but without a good justification for sharing it. If we accept that confidential data that is also useful carries with it an inherent risk of re-identification, the role of the data custodian then becomes one of ensuring that the risk is **negligible**.

We can think of the ‘anonymised’ concept in ‘anonymised data’ in the same way we think of the ‘reinforced’ concept in ‘reinforced concrete’. We do not expect reinforced concrete to be indestructible, but we do expect that a structure made out of the stuff will have a negligible risk of collapsing (Elliot, Mackey and O’Hara, 2016).

A risk-based approach to anonymisation is implemented in practice by National Statistical Institutes around the world. For example, the UK’s Office for National Statistics carries out an enormous amount of work in the area of scenario modelling, statistical disclosure risk assessment and management, in order to provide useful confidential census data in varying formats to a wide range of audiences. This is also a position supported by the UK’s Information Commissioner’s Office, as described in its 2012 Code of Practice on Anonymisation:

**‘The DPA does not require anonymisation to be completely risk free, you must be able to mitigate the risk of identification until it is remote. If the risk of identification is reasonably likely the information should be regarded as personal data, these tests have been confirmed in binding case law from the High Court’**(2012:6)**.”**

# How do you do anonymisation?

The two most common approaches to doing anonymisation are: data centric anonymisation and functional anonymisation. We will now look at each of these in turn.

## Data centric anonymisation

Data centric anonymisation is the approach most commonly taken. It essentially views re-identification risk as originating from and contained within the data to be released. It is an approach that asks: ‘how risky is the data to be shared or published’? What this means in practice is that re-identification risk is assessed and managed largely by taking account of the data only. Little or no attention is given to other wider and key considerations such as how, or why, a re-identification might happen, or what skills, knowledge, or other data a person would require to ensure their attempt was a success. Risk doesn’t arise from data per se but from the interaction between data and its environment (including who is looking at it). The data centric approach undoubtedly underpinned the release of the hospital insurance dataset by GIC that led to the re-identification of Governor, and to the Netflix case.

## Functional anonymisation

Functional anonymisation (FA) is a more recent approach to doing anonymisation. FA essentially shifts the focus from data to a focus on the relationship between data and the environment in which it exists to assess and manage re-identification risk. It is an approach that shifts the question of ‘**how risky is the data’ to a question of ‘how risky is the proposed share or publication given the data and the share/publication environment’.**

So what do we mean by the term ‘data environment’?

The data environment is best imagined as being made up of four component features (Mackey & Elliot 2013; Elliot & Mackey 2014):

* **Other data:**any information that could potentially be linked to the data in question thereby enabling re-identification. There are four key types of other data: personal knowledge, publicly available sources, restricted access data sources, and other similar data releases.
* **Agents:** those people and entities capable of acting on the data and interacting with it along any point in a data flow.
* **Presence or absence of Governance Processes:**processes that essential inform the **who and how of access** to data. They include formal governance e.g. data access controls, licensing arrangements and policies which prescribe and proscribe agents’ interactions and behaviour through norms and practices.
* **Presence or absence of Infrastructure:** i.e. the set of interconnecting structures (physical, technical) and processes (organisational, managerial) that frame and shape the data environment.

In practice, thinking about the relationship between data and the component features of the data environment can help you to assess risk and in turn mitigate it.

To illustrate, let us consider two simple examples — one where data is published in an open environment (internet) and one where data is shared in a controlled and restricted environment (Trusted Research Environment).

## **Open environment: publication on the internet**

An open environment such as publication on the internet, we can imagine is configured in the following way:

* In terms of agents that could potentially access the published data, the pool is anyone in the world with access to the internet.
* In terms of other data, it is all other data that co-exist on the internet as well as data that co-exist off-line that could potentially be linked to the data to be published.
* In terms of governance processes and infrastructure, there is likely to be a near total absence of both components in respect to restrictions or controls on the who and how of access.

In an open environment then, to manage risk and achieve functional anonymisation, you will need to restrict, quite considerably, the data to be released. Even aggregate data such as tables and graphs can be potentially disclosive.

For information on how to ensure your research outputs intended for publication are not disclosive please see:Bond et al, Guidelines for the checking of output based on microdata research and Griffiths et al, July 2019, Handbook on Statistical Disclosure Control (SDC) for Outputs.

## **Secure environment: Trusted Research Environment**

A secure environment such as a Trusted Research Environment (TRE) is typically configured in the following way:

* In terms of**agents** who can have access to the data –typically TRE’s operate an application and approvals process to vet both researcher and the project (i.e. intended use of the data) and require the researcher to complete Researcher Accreditation Training and sign a Terms and Conditions of Access Agreement
* In terms of **what other data** could be linked with the data in the TRE — typically TRE’s operate strict protocols on what can be brought into and out of the TRE, including such as prohibiting mobile phones and pen and paper being brought in and checking statistical results (intended to be brought out) to ensure that they are non-disclosive
* In terms of **governance processes and infrastructure**— TRE’s operate under strict IT security protocols and may also operate secure physical facilitates. This may include a secure locked facility, monitoring of people entering and leaving the facility, secure IT infrastructure monitoring access to data and prohibiting the movement of data within the data store without explicit permissions.

In a secure controlled environment then, risk is being tightly managed and mitigated. It may be possible to class the data shared (even detail data) as functionally anonymised for the researcher. For more information on this please see Elliot et al 2016, 2020, Mourby et al 2018.

# A Framework for doing anonymisation

To support well thought out anonymisation the UK Anonymisation Network (UKAN) published the Anonymisation Decision Making Framework (ADF) book in 2016 and most recently a new edition in the form of a [European Practitioner’s Guide](https://ukanon.net/framework/). It is a practical guide to GDPR-Complaint anonymisation that provides more operational advice than most other publications.

The ADF is primarily intended to support practitioners to anonymise with confidence. It comes with tools and templates to capture and evaluate different use cases.

The ADF is underpinned by four core principles:

## **Comprehensiveness Principle**

**You cannot decide whether or not data are safe to share/release by looking at the data alone, but you still need to look at the data.**

This principle encapsulates the data situation approach where risk is seen as arising from the interaction between data, people, other data, governance processes and infrastructure. You do also need to know your data — which means being able to identify the critical properties of your data and to assess how they might affect risk. This will feed into decisions about how much data to share or release, with whom and how.

## **Utility Principle:**

**Anonymisation is a process to produce safe data, but it only makes sense if what you are producing is safe useful data.**

Anonymisation is not just about data protection but also data utility.

## **Realistic Risk Principle:**

**Zero risk is not a realistic possibility if you are to produce useful data.**

This is fundamental. Functional anonymisation is about risk management. Accepting that there is a residual risk in all useful data inevitably puts you in the realms of balancing risk and utility. However, the trade-off of individual and societal benefits against individual and societal risks is the stuff of modern life.

## **Proportionality Principle**

**The measures you put in place to manage risk should be proportional to that risk and its likely impact.**

Following the realistic risk principle, the existence of risk is not necessarily a reason for withholding access to data. However, a mature understanding of that risk will enable you to make proportionate decisions about the data, who should have access and under what conditions.

The ADF comprises of three core anonymisation activities and 10 component features, they are:

## **Activity 1: Data situation audit**

This is essentially a framing tool for understanding your data situation. The audit will help you effectively assess and manage risk. There are 6 components associated with this anonymisation activity; they are:

1. Describe/capture the presenting problem
* A top-level description of what you are trying to achieve or do.

2. Sketch the data flow and determine your responsibilities

* This will enable you to visualise the outline of your data situation and identify your responsibilities within it.

3. Map the properties of the data environment(s) [DE]

* Once the data flow is sketched you can map on the four components of the environment(s) i.e. other data, agents, governance and infrastructure.

4. Describe and map the data

* Once you’ve you mapped the properties of your DE, you need to describe the data within each environment across a range of parameters i.e. data structure, data type, variable type, population, dataset properties etc.

5. Engage with stakeholders

* This is crucial to trustworthy data stewardship

6. Evaluate the data situation

* An evaluation of 1–5, will inform whether you can proceed to share /publish the data or whether you need to assess risk in more detail and put further controls on that risk.

## **Activity 2: Risk analysis control**

Once you have completed an evaluation of your data situation (component 6) you will be in a position to be able to decide whether further risk assessment and control is necessary. How to go about assessing risk and implementing control methods are detailed in component 7. Risk assessment and control should usually be an iterative, not linear, process. You may need the advice of an expert at this stage.

7. Select and implement the processes you will use to assess and control disclosure risk.

## **Activity 3: Impact management**

Impact management is about making plans to manage an adverse event if in the rare event one should happen. There are three components associated with this anonymisation activity, they are:

8. Maintain stakeholders’ trust.

9. Plan what to do if things go wrong.
• Residual risk means that an adverse event could happen so it is important to have in place a crisis management policy.

10. Monitor the Data Situation.
• Risk is neither exactly calculated nor constant. You should produce and implement a policy for monitoring the risk and consider adjusting the data situation if it changes significantly.

# Summary

In this resource we explored how anonymisation and pseudonymisation applies to the management, sharing and publication of research data.