# Chapter 2: How do we ensure anonymisation is effective?

## Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance

October 2021

ico.
Information Commissioner's Office

# Contents

# How do we ensure anonymisation is effective?

## At a glance

- Effective anonymisation reduces identifiability risk to a sufficiently remote level.
- Identifiability is about whether someone is “identified or identifiable”. This doesn’t just concern someone’s name, but other information and factors that can distinguish them from someone else.
- Identifiability exists on a spectrum, where the status of information can change depending on the circumstances of its processing.
- When assessing whether someone is identifiable, you need to take account of the “means reasonably likely to be used”. You should base this on objective factors such as the costs and time required to identify, the available technologies, and the state of technological development over time.
- However, you do not need to take into account any purely hypothetical or theoretical chance of identifiability. The key is what is reasonably likely relative to the circumstances, not what is conceivably likely in absolute.
- You also need to consider both the information itself as well as the environment in which it is processed. This will be impacted by the type of data release (to the public, to a defined group, etc)  and the status of the information in the other party’s hands.
- When considering releasing anonymous information to the world at large, you may have to implement more robust techniques to achieve effective anonymisation than when releasing to particular groups or individual organisations.
- There are likely to be many borderline cases where you need to use careful judgement based on the specific circumstances of the case.
- Applying a “motivated intruder” test is a good starting point to consider identifiability risk.
- You should review your risk assessments and decision-making processes at appropriate intervals. The appropriate time for, and frequency of, any reviews depends on the circumstances.

## In detail

- What should our anonymisation process seek to achieve?
- What is identifiability?
- What are the key indicators of identifiability?

## What should our anonymisation process seek to achieve?

An effective anonymisation process seeks to reduce the likelihood of someone being identified or identifiable to a sufficiently remote level. This level depends on a number of factors specific to the context.

It may seem fairly easy to say whether a piece of information relates to an **identified** individual, as this may be clear from the information itself. For example, bank statements clearly identify individual account holders and contain information that relates to them.

It may seem less clear whether someone is **identifiable**. However, it is important to note that data protection law defines personal data as:

> **Quote**
>
> 'any information relating to an **identified or identifiable** living individual'

The law also says that an "identifiable living individual" is someone:

> **Quote**
>
> '…who can be identified, directly or indirectly, in particular by reference to:
>
> (a) an identifier such as a name, an identification number, location data or an online identifier; or
>
> (b) one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the individual.'

Therefore, anonymisation processes should take into account the concept of identifiability in its broadest sense. They should not simply focus on removing obvious information that clearly relates to someone.

## What is identifiability?

In its basic meaning, identifiability is about whether someone is identified or identifiable. Understanding this concept is crucial to ensure you then understand the nature of the information you hold.

Essentially, if you can distinguish an individual from other individuals, then they are identified or identifiable. This is also known as singling out or individuation.

While a name may be the most common way someone is identifiable, it is important to understand that:

- an individual can be identifiable even if you do not know their name. If there is other information that enables individuals to be connected to data that could only be about them, then they may still be identified or identifiable; and
- whether any potential identifier actually means an individual is identifiable depends on the context.

Identifiers are pieces of information that can be closely connected to particular individuals, and include:

- direct identifiers (eg someone's name); and
- indirect identifiers (eg a unique identifier you assign to them such as a number).

Data protection law provides a non-exhaustive list of common identifiers in the definition of personal data itself. For example, name, identification number, location data and online identifier. However, as detailed above, the definition also specifies other factors such that can mean an individual is identifiable..

This means that simply removing direct identifiers from a dataset is insufficient to ensure effective anonymisation. If it is possible to link any individuals to information in the dataset that relates to them, then the data is personal data.

You also need to consider data other than identifiers and how it may be used to provide context that can single out an individual. For example, images or information about someone's location.

At the same time, the existence of identifiers does not always mean that individuals are identified or identifiable. Contextual factors are also important.

For example, information about an individual's year of birth may allow them to be singled out in the context of their family, but not in the context of a different group like their class at school. Similarly, someone's family name may be enough to distinguish them from others in the context of their workplace, but not in the context of the general population (eg Smith or Jones).

It is important to note that data that may appear to be stripped of identifiers can still be personal data in cases where it can be combined with other information and linked to an individual. For example, data available publicly, or to a particular organisation. Even if stripping identifiers is not sufficient to achieve anonymisation, doing so may still be a sensible approach in the context of the data minimisation principle, eg if such identifiers are not required.

In most cases, a unique identifier will mean you can distinguish someone from someone else. For example, an NHS number is different for every individual and therefore will allow them to be singled out from other individuals in the dataset.

**Relevant provisions in the legislation**

Section 3 of the DPA 2018 (external link)

UK GDPR Article 4(1) and Recital 30 (external link)

**Further reading – ICO guidance**

See our guidance on 'What is personal data' for more information about:

identifiers and related factors;

direct identification; and

indirect identification.

## What are the key indicators of identifiability?

Reducing identifiability to sufficiently remote levels can seem challenging given the broad definition of personal data. In the context of the information you hold and the end goal of your anonymisation process, it may be useful for you to consider three key indicators for determining whether information is personal data or not. These are:

- singling out;
- linkability; and
- inferences.

Effective anonymisation techniques seek to reduce the likelihood of these three occurring.

### What is ‘singling out’?

You need to consider whether singling out is possible, both by you and by another party. This should be part of your assessment of the effectiveness of your anonymisation processes.

The general processing regime in the UK GDPR specifically references singling out as something you need to address when you consider the concept of identifiability.

As noted above, singling out means that you are able to tell one individual from another individual in a dataset. For example, if you can isolate some or all records about an individual in the data you process, then that individual is singled out.

It is important to note that even if you do not intend to take action about an individual, the fact that they can be singled out may allow you to do so. They are therefore still identifiable.

To determine the possibility of singling out, you need to consider the richness of the data and how potentially identifying different categories are. You also need to consider whether sufficient safeguards are in place to reduce this risk.

### What is ‘linkability’?

Linkability is the concept of combining multiple records about the same individual or group of individuals together. These records may be in a single system or across different systems (eg within one database, or in two or more different databases).

Linkability is sometimes known as the mosaic or jigsaw effect. This is where individual data sources may seem non-identifying in isolation, but can lead to the identification of an individual if combined.

Common techniques to mitigate linkability include masking and tokenisation of seemingly identifying key variables. For example, sex, age, occupation, place of residence, country of birth.

Linkability is also a crucial consideration for pseudonymisation. The existence of additional records that could be linked may be regarded as ‘additional information’ that enables identifiability. In turn, this means the information in question is personal data that has undergone pseudonymisation, rather than anonymous information.

### What are ‘inferences?’

An inference refers to the potential to **infer, guess** or **predict** details about someone. In other words, using information from various sources to deduce something about an individual (eg based on the qualities of others who appear similar).

Inferences may also be the result of analytical processes intended to find correlations between datasets, and to use these to categorise, profile, or make predictions about people.

An inference can therefore be something you create, as opposed to something that you collect or observe.

Whether an inference is personal data depends on whether it relates to an identified or identifiable individual.

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

- incomplete datasets, eg, where some of the identifying information has been removed or generalised;
- from pieces of information in the same dataset that are not obviously or directly linked; or
- 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.

You should also consider whether the specific knowledge of others, such as doctors, family members, friends and colleagues could be sufficient additional information that may allow inferences to be drawn.

## What is the "spectrum of identifiability"?

In one sense, data protection law presents a simple binary outcome – information either meets the definition of personal data or it does not.

At the same time, the actual identifiability of individuals in practice can be highly context-specific. Different types of information have different levels of identifiability risk depending on the circumstances in which you, or another party, process them.

Whether something is personal data or anonymous information is therefore an **outcome** of assessing identifiability risk, taking into account the relevant facts.

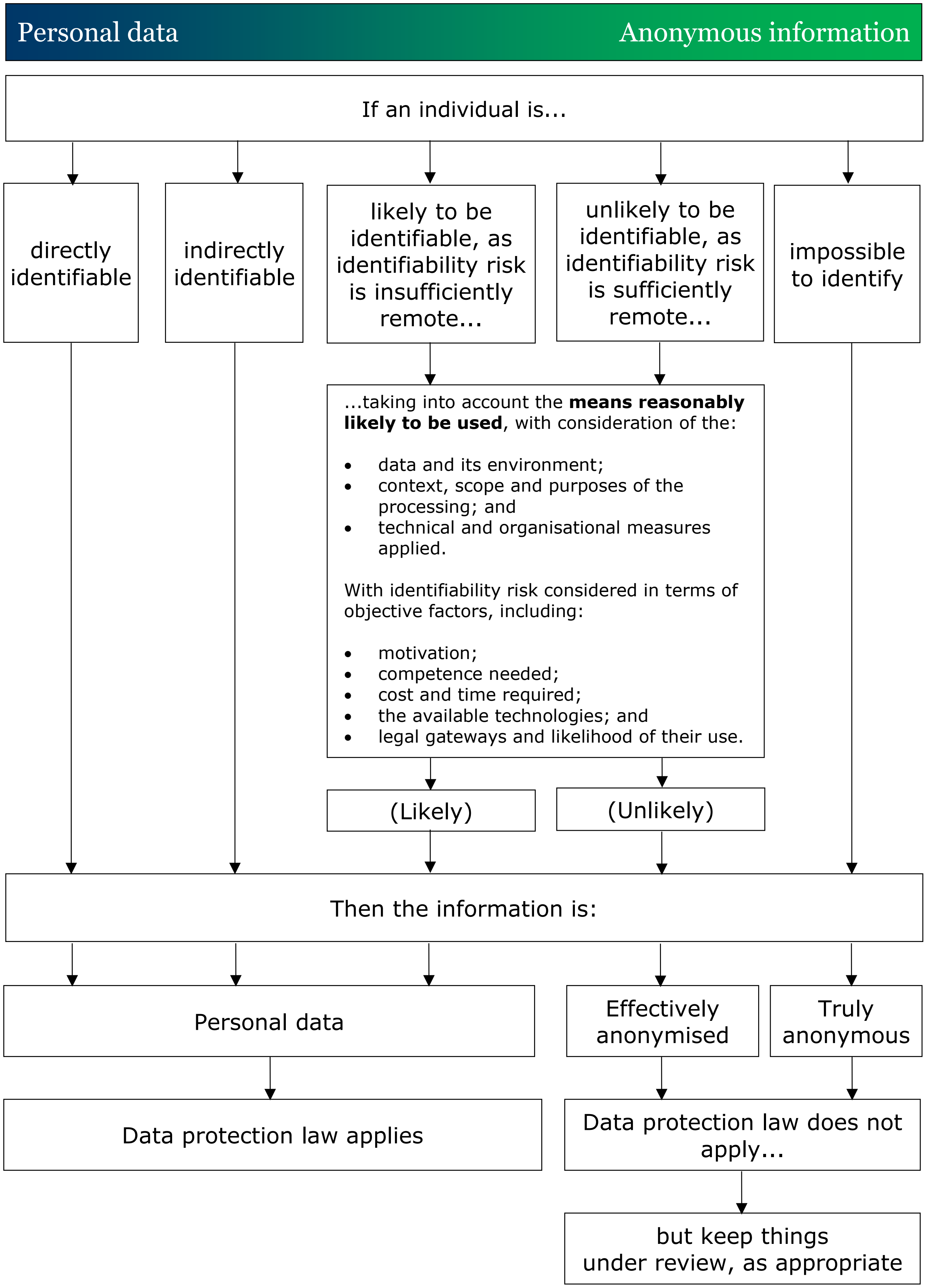
In practice, identifiability may be viewed as a spectrum that includes the binary outcomes at either end, with a blurred band in between. For example:

- at one end, information relates to directly identified or identifiable individuals (and will always be **personal data**); and
- at the other end, it is impossible to relate information to an identified or identifiable individual. This is **anonymous information**.

For everything in between, identifiability depends on the specific circumstances and risks posed. Essentially, information may 'move' along the spectrum of identifiability to the point that data protection law starts to apply to it (or, conversely, stops applying to it).

There are a number of ways that the spectrum of identifiability may be visualised. These may be specific to certain industries or sectors, or may relate to particular practices.

In this guidance, we do not intend to endorse or prohibit approaches that work for particular organisations or industries in the UK. What is important is ensuring that any approach you take considers the requirements of data protection law. We provide one way of considering the spectrum in this context below.



Personal data

Anonymous information

If an individual is...

directly identifiable

indirectly identifiable

likely to be identifiable, as identifiability risk is insufficiently remote...

unlikely to be identifiable, as identifiability risk is sufficiently remote...

impossible to identify

...taking into account the **means reasonably likely to be used**, with consideration of the:

- data and its environment;
- context, scope and purposes of the processing; and
- technical and organisational measures applied.

With identifiability risk considered in terms of objective factors, including:

- motivation;
- competence needed;
- cost and time required;
- the available technologies; and
- legal gateways and likelihood of their use.

(Likely)

(Unlikely)

Then the information is:

Personal data

Effectively anonymised

Truly anonymous

Data protection law applies

Data protection law does not apply...

but keep things under review, as appropriate

**Figure 1: Mapping the concept of the spectrum of identifiability to data protection law**

Information may shift towards one end of the spectrum, depending on factors including:

- the specifics of the processing. For example, the sensitivity of the variables in the original dataset and the techniques you use to reduce the identifiability of individuals in the data;
- the data environments involved. For example, the technical and organisational measures in place to control access to the data and reduce identifiability risk, and
- your risk management process. For example, how you identify and mitigate any risks of the processing.

This means the status of information – as personal data or anonymous – can change over time.

**Further reading outside this guidance**

Some examples of how the concept of the spectrum of identifiability can be visualised include:

- Understanding Patient Data's 'Identifiability Demystified' briefing (external link, PDF);
- the Future of Privacy Forum's 'Visual guide to practical data de-identification' (external link, PDF);
- the National Institute of Standards and Technology (NIST) publication 'De-Identification of Personal Information' (NISTIR 8053) (external link, PDF); and
- Privacy Analytics' presentation 'Principles of de-identification' (external link, PDF).

These examples are to illustrate different approaches. Their reference here does not represent an ICO endorsement.

## What does data protection law say about assessing identifiability risk?

When assessing identifiability risk the core question you need to ask is whether there are "means reasonably likely to be used" to identify an individual.

The general processing regime in the UK GDPR provides additional information about the factors you need to take into account when determining this. Similar considerations apply to Parts 3 and 4 of the DPA 2018.

Recital 26 of the UK GDPR states that:

**Quote**

'To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'

This means that, once you take all objective factors into account:

- if there are means "reasonably likely" to be used to identify someone, then you must view the information as personal data; and
- if no means are "reasonably likely" to be used, then you can view the information as effectively anonymised. However, the identifiability risk must be sufficiently remote in the context of the processing.

As noted in earlier sections of this guidance, information that is effectively anonymised is not personal data and data protection law does not apply.

The feasibility and cost-effectiveness of the available means to identify individuals can change over time. In general, the more feasible and cost-effective a method becomes, the more likely it is to be a means reasonably likely to be used.

The measures reasonably likely to be taken to identify an individual may vary depending upon the perceived value of the information. For example, if the information is thought to be about a high profile public figure, it is likely that there will be some who are willing to use more complex measures to identify that individual. In this context, these would still be "means reasonably likely to be used", even if there may not be in other cases.

**Relevant provisions in the legislation**

UK GDPR Article 4(1) and Recital 26 (external link)

## How should we approach this assessment?

You should consider the means reasonably likely to be used at the earliest stage of your anonymisation process, particularly when deciding the "release model" (ie public release, release to defined groups etc). In all release cases, your assessment of identifiability requires you to consider:

- whether there is additional information that may enable identification;
- whether there are techniques that enable identification from the information in question; and
- the extent to which the additional information or techniques are reasonably likely to be accessible to (and used by) a particular person to identify individuals the original information relates to.

In some cases, the risk of anonymous information being combined with other data to result in personal data being created may be high. For example, where:

- anonymised data can be combined with publicly-available information meaning someone becomes identifiable; or
- complex statistical methods may 'piece together' various pieces of information with the same result.

It is important to note that different additional information or techniques may be available to different parties, depending on the circumstances. This means that the status of the information may change. For example, the same information may:

- be personal data in your hands. For example, if you also hold additional information that means individuals are identifiable (even if you hold this separately and apply technical and organisational controls to it); and
- not be personal data once in the hands of other parties. For example a specific recipient (or the general public), if they have no access to the additional information and no means reasonably likely to be used to obtain it.

Additionally, even if you determine that the data you hold does not allow the identification of individuals today, that position may change in the future. For example, due to new technologies or developments or changes to the public availability of certain records.

Data protection law does not require you to adopt an approach that takes account of every absolute or purely hypothetical or theoretical chance of identifiability. 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.

The key is what is 'reasonably likely' relative to the circumstances, not what may be 'conceivably likely' in absolute.

Effective anonymisation is about finding the right balance between managing this risk while keeping the utility of the data. It may not be possible to determine with absolute certainty that no individual will ever be identifiable as a result of the disclosure of anonymous information. However, you can adopt a certain amount of pragmatism.

## What factors should we include?

You should approach assessing identifiability risk by considering what is reasonably likely relative to the context. This includes whether identification is technically and legally possible, taking into account objective criteria including:

- how costly identification is in human and economic terms;
- the time required for identification; and
- the state of technological development at the time of processing (ie the techniques you use anonymising the data, and/or when you are sharing the dataset with another party); and
- future technological developments (ie as technology changes over time).

You also need to frame this assessment in the context of the specific risks that different types of data release present. For example, these can differ if you are disclosing information to:

- another organisation;
- a pre-defined group of organisations; or
- the wider public or the world at large.

Different scenarios present different challenges and potential harms that you need to mitigate. For example, when you disclose or otherwise make available information to another organisation, both you and they should assess identifiability risk. You need to clearly establish the status the information has in your respective hands.

The greater the likelihood that someone may attempt to identify an individual from within a dataset, the more care you have to take to ensure effective anonymisation. However, as noted above, you do not need to take account of every hypothetical or theoretical risk of identifiability. The key is whether identifiability is ‘reasonably likely’ given the circumstances.

To determine what is reasonably likely, you need to decide what level of identifiability risk is acceptable. Assessment of this risk is contextual. It requires you to consider:

- the information itself;
- its environment, including the restrictions placed upon the data sharing, the sensitivity of the data, the potential linkage of released data with other data; and
- a re-examination of the robustness of anonymisation to consider new technologies and threats, as appropriate.

The cost, time taken and technology required for identification are impacted by the nature of the data environment, including technical and organisational measures and contractual controls it has. This can limit the ability of any attacker to identify people and thus further reduce the risk.

For example, you may apply techniques to the data, such as generalisation and randomisation, which transform it so that identifiability risk is reduced. However, there may be other circumstances where you may not be able to apply any controls to the environment, such as with open data release.

You also need to take account of how this risk may change as information moves from one environment to another, depending on what is shared and the controls put in place.

Other factors also impact the environment, including:

- additional data that may exist (eg other databases, personal knowledge, publicly available sources);
- who is involved in the processing, and how they interact;
- the governance processes that are in place to control how the information is managed (eg who has access to it and for what purposes); and
- the legal considerations that may apply, such as:
    - any gateways that may impact the potential for disclosing information that enables individuals to be identifiable; or
    - prohibitions that mean while information could technically be combined to aid identifiability, doing so is against the law (eg professional confidentiality).

If, taking the above into account, you conclude that the likelihood of identifiability is sufficiently remote then your assessment may be that the information is effectively anonymised. However, you need to:

- document and justify your decision; and
- keep this under review (eg as technologies change over time).

A good starting point for your assessment is to consider the concept of the "motivated intruder", and including appropriate tests in your decision-making and review processes.

## What is the "motivated intruder" test?

Personal data is not just important to the individuals it relates to, but to others that may be motivated to obtain it. However, data protection law does not specify how you determine whether:

- the anonymous information you release is likely to result in the identification of an individual; or

- anyone has the motivation to carry out that identification.

A useful test to include as part of assessing identifiability risk is whether an intruder would be able to achieve identification if they were motivated to attempt it. This is known as the motivated intruder test. It is used by both the ICO and the Information Tribunal, which hears DPA 2018 and FOIA appeals.

The test is useful because it sets the bar for assessing the risk:

- higher than simply considering whether a 'relatively inexpert' member of the public can achieve it; but
- lower than considering whether someone with access to significant specialist expertise, analytical power or prior knowledge could do so.

We recommend that you adopt a motivated intruder test as part of your risk assessment. It is also good practice to use the test as part of any review, both of your overall risk assessment and the techniques you use to achieve effective anonymisation.

**Who is a motivated intruder?**

A motivated intruder is a person who starts without any prior knowledge but wishes to identify an individual from whose personal data the anonymous information is derived. The test assesses whether the motivated intruder is likely to be successful.

It assumes that a motivated intruder is someone that:

- is reasonably competent;
- has access to appropriate resources (eg the internet, libraries, public documents); and
- uses investigative techniques (eg making enquiries of people who may have additional knowledge about an individual, or advertising for anyone with that knowledge to come forward).

The intruder is therefore someone who has the:

- motives to attempt identification;
- means to succeed; and
- intent to use the data in ways that may pose risks to your organisation and the rights and freedoms of individuals whose data you process.

You should assume that you are not looking just at the means reasonably likely to be used by an ordinary person, but also by a determined person with a particular reason to want to identify individuals. For example, intruders could be investigative journalists, estranged partners, stalkers, or industrial spies.

As a baseline, a motivated intruder is not assumed to have:

- specialist knowledge (eg in-depth knowledge of computer hacking);
- access to specialist equipment; or
- the need to resort to criminal acts to gain access to data that is held securely (eg burglary).

At the same time, different types of potential attacker and different motivations may mean that the profile of a likely intruder also differs. For financial data, confidential files and other types of high-value data you must also consider intruders with stronger capabilities, tools and resources.

For example, state actors may have access to significant computing power and expertise. Whether you need to factor this into your identifiability risk assessments depends on your circumstances. However, it may require you to implement stronger technical and organisational measures to mitigate the additional risks when compared to ”business as usual” processing operations.

The intruder can be someone who is not intended to have access to the information, as well as someone who is permitted this access but may identify an individual, intentionally or accidentally.

In essence, your motivated intruder test should consider:

- the nature, type and volume of information you process;
- the likelihood of someone wanting to attempt to identify individuals, for whatever purpose;
- the range of capabilities an intruder may have;
- the information that they may already have (or can access); and
- the controls you deploy within your data environment to prevent this.

**What types of motivations are there?**

A motivated intruder can be classified in several ways, depending on their status and background knowledge. For example, you should consider:

- their relationship to an individual;
- their background knowledge;
- whether they are targeting a specific or random individual(s) in the dataset
- whether they know (with a degree of certainty) that the individual is in the data set; and
- their access to specialist resources and expertise.

Clearly, some types of data will be more attractive to a motivated intruder than others. Obvious motivations may include:

- finding out personal data about someone else, for nefarious reasons or financial gain;
- the possibility of causing mischief by embarrassing others, or to undermine the public support for release of data;
- revealing newsworthy information about public figures;
- political or activistic purposes (eg as part of a campaign against a particular organisation or person);
- curiosity (eg a local person’s desire to find out who has been involved in an incident shown on a crime map);
- a demonstration attack in which a hacker or researcher is interested in showing that identification of individual(s) is possible; or
- a random inadvertent recognition of an individual by a well-known acquaintance.

This does not mean that you can simply release data which is seemingly ordinary, innocuous or otherwise without value. You still need to undertake a thorough assessment of identifiability risk to determine the potential impact on individuals.

**Example**

With health data, there may be no obvious motivation for trying to identify the individual that a particular patient ‘episode’ relates to. However, the degree of embarrassment or anxiety that re-identification could cause may be very high.

The anonymisation techniques employed to protect this data need to reflect these potential harms.

**How does the type of data release matter?**

There is a clear difference between releasing data to the world at large and making it available to a smaller defined group.

With public release, it may be virtually impossible for you to retract the data if it later becomes clear that identifiability is reasonably likely. You also do not have control over the actions and intentions of any recipients of that information. These factors may pose more challenges than other contexts. In these circumstances, your approach to anonymisation needs to be very robust in order to be effective.

With release to defined groups, your identifiability risk assessment should consider the information and technical know-how available to members of that group. Contractual arrangements (eg binding restrictions) and associated technical and organisation controls play a role in the overall assessment. Fewer challenges may arise than with public release.

This is particularly the case if you retain control over who can access the data, and the conditions in which they can do so. Designing these access controls appropriately will help reduce identifiability risk and potentially allow you to include more detail, while continuing to ensure effective anonymisation.

You do still need to consider the possibility that the data may be accessed by an intruder from outside the group, or that it may be shared inappropriately. You should address this with physical and technical security controls aimed at preventing this access. If there is a greater likelihood of accidental release or unauthorised access, your identifiability assessment needs to demonstrate how you intend to mitigate this risk.

As part of your identifiability risk assessment, you should therefore consider the circumstances of the data release. For example:

- in cases of public disclosure or open data release, you should consider the maximum risk of re-identification across all records in the dataset; and
- in non-public data release scenarios, you should consider contractual controls and limitations on how the data is accessed, used and disposed of. These should be supported by technical and organisational measures.

This can also impact how you apply the motivated intruder test, as different attacks and motivations can apply depending on the nature of the release.

## What information can a motivated intruder use?

When considering the motivated intruder test it is useful to think about the different types of information that may be available. For example, information that an intruder may:

- possess, eg background or prior knowledge; and
- learn, eg by searching publicly-available sources, etc.

**What ground or prior knowledge could an intruder possess?**

Background and prior knowledge depend on the relationship between the intruder and the individual they wish to identify. You should consider the following factors:

- the likelihood of individuals having and using the knowledge to allow identification; and
- the likely consequences of this identification, if any.

Identifiability risk can arise where one individual or group knows a great deal about another individual. They may be able to determine that 'anonymised' data relates to a particular individual, even if an ordinary member of the public would be unable to. For example:

- a doctor could determine that an anonymised case study in a medical journal relates to a patient they have treated;
- one family member may work out that an indicator on a crime map relates to an incident involving another family member; and
- an employee may work out that a particular absence statistic relates to a colleague who they know is on long-term sick leave.

This is an example of why identifiability risk is contextual, and may be unable to be rule out entirely. Those with particular personal knowledge might learn something about another individual, even if this only confirms an existing suspicion.

However, the risks may be lower in other cases, and a relevant factor is whether someone would learn anything new. For example, whether an individual recognises that anonymous information relates to them, allowing self-identification to take place.

You should not make assumptions about family relationships and what individuals may already know. For example, teenagers may not share certain medical information with parents or other family members.

The likelihood of identifiability may be difficult to assess in the context of large datasets or collections of information. In these cases it is more practical to consider a more general assessment of the risk of prior knowledge leading to identification. For example, for identification of at least some individuals recorded in the information. You could then make a global decision about the chances that those who might be able to re-identify are likely to seek out or come across the relevant data.

The likely consequences can also be difficult to assess in practice. A member of the public's sensitivity may differ from yours. For example, the disclosure of the address of a person in a witness protection scheme could be more consequential than in other cases.

It is reasonable to conclude that certain professionals with prior knowledge, are not likely to be motivated intruders (eg doctors). This could apply where it is clear that the profession in question imposes confidentiality rules and requires ethical conduct.

It is also good practice to consult and understand the experience of other groups. Something that you might not feel needs to be protected could have dramatic negative consequences for people in different circumstances.

### What is the difference between information, established fact and knowledge?

When you consider these issues, it is also useful to distinguish between recorded information, established fact, and knowledge.

**Example**

- "Mr B. Stevens lives at 46 Sandwich Avenue, Stevenham."

  This may be established fact (eg because the information is contained in an up-to-date copy of the electoral register).

- "I know Mr B. Stevens is currently in hospital, because my neighbour – Mr Stevens' wife - told me so."

  This may be personal knowledge, because it is something that Mr Stevens' neighbour knows.

The starting point should be to consider recorded information and established fact. It is easier to establish that particular information is available than to work out whether an individual has the knowledge necessary to allow for identification.

It is still the case that non-recorded personal knowledge, in combination with anonymous information, can lead to identification. However, in practice there must be a plausible and reasonable basis for this to be considered in order for it to present significant identifiability risk.

### What about educated guesses?

Identifiability involves more than making an educated guess that information is about someone. Data protection law concerns information that identifies someone, which implies a degree of certainty that information is about one person and not another.

As described above, if the information allows someone to be distinguished or singled out from someone else, this will usually be sufficient.

However, the mere possibility of making an educated guess about whether an individual is identifiable does not necessarily present a data protection risk. Even where a guess based on anonymous information is correct, this does not mean that a disclosure of personal data has happened.

At the same time, the consequences of releasing the anonymous information may be such that you should adopt a cautious approach, even where a disclosure would not be a disclosure of personal data.

This is a complex area and when approaching these issues it can be helpful to look primarily at the possible impact on individuals, and then whether or not there is likely to be a disclosure of personal data.

**What can an intruder learn about an individual?**

At a minimum, assume a motivated intruder is someone who gathers information on particular individuals by extensive searching of internet sources, possibly including some low-cost subscription services. More determined intruders may be willing to incur additional costs and take extra steps.

If you consider the typical steps and types of information, you can begin to identify what an intruder is likely to learn about individuals. help you identify what an intruder can learn about individuals. In turn, this can enable you to carry out the motivated intruder test in practice.

Obvious sources of information include:

- libraries;
- local council offices;
- church records;
- public records (eg General Register Office, the electoral roll, the Land Registry);
- genealogy websites;
- online services (eg social media, internet searches);
- local and national press archives; and
- releases of anonymous information by other organisations (eg public authorities).

Intruders with criminal intent may use illegal means to gather potentially matching data (eg fraud). For example, creating false accounts, carrying out social engineering attacks or impersonation to gain further information that could be used for re-identification.

Limiting access to data where possible and close consideration of the safeguards that you can adopt can help to reduce the risk in this case.

## Do we need to consider who else may be able to identify individuals from the data?

Yes. You also need to consider whether it is reasonably likely that someone else can identify individuals. This could either be just from the information in question or from that and other information they may possess or obtain.

These considerations are particularly relevant where you apply a technique to personal data and intend to make the resulting dataset accessible to another party. For example, by sharing it with them or storing it in an environment whose access you control.

This can sometimes be known as the 'whose hands?' question (ie what is the status of the information in the respective 'hands' of those who process it?).

Overall, your anonymisation processes need to take account of the nature, scope, context and purposes of the processing, as well as the risks it poses. These are likely to differ from one organisation to another and from one context to another. While there may be circumstances where these considerations are similar, in general you cannot apply one single formula that will guarantee effective anonymisation in all instances.

**Example: Disclosure between organisations**

Organisation A creates a dataset by treating personal data in such a manner that identifiability risk is subject to particular controls. Organisation A intends to disclose the resulting dataset to Organisation B.

In the hands of Organisation A, the data is likely to remain in scope of data protection law. For example, where it keeps the original personal data, or possesses the additional information that enables re-identification (even if held separately and subject to particular technical and organisational measures). Identifiability is reasonably likely (or already present) in both cases.

In the hands of Organisation B, the dataset's status may be different, depending on the circumstances. The key factor is for both parties to assess this status in the context of the disclosure.

For example, prior to the disclosure, Organisations A and B should assess identifiability risk based on the objective criteria this guidance describes, taking into account the circumstances in which:

- Organisation A creates the dataset; and
- Organisation B intends to process the dataset.

It may be appropriate, and indeed more practical, for both organisations to undertake this assessment jointly.

The result of this assessment may be that in the hands of Organisation B the identifiability risk is either:

- sufficiently remote, so from its perspective the information is effectively anonymised; or
- insufficiently remote, so from its perspective the information is personal data.

Both parties should document the outcome of this assessment.

These considerations are still relevant to the decision to disclose the data, even though in practice Organisation A is likely to have limited or no control of:

- the data environment of Organisation B; or
- the overall circumstances in which the information is processed once in the hands of Organisation B.

This is particularly the case if the dataset is subject to unauthorised or unlawful re-identification when in the hands of Organisation B (eg at a future point in time). However, for Organisation A, the key is that its decision to release the data resulted from a rational thought process that took identifiability risk into account.

Where the information disclosed is personal data, Organisations A and B are essentially entering into a data sharing arrangement. In this case, both organisations should instead consider the requirements of the ICO's data sharing code of practice.

The origin of a dataset may form a factor in any investigation we may undertake. However, this does not automatically mean the disclosing organisation will be at fault. Any regulatory action we take will depend on the specific circumstances of the case.

We cannot fully rule out the possibility that the disclosing organisation may also have some responsibility in any such breach. However, having a documented and justifiable assessment of identifiability risk will help you demonstrate that your approach to anonymisation is effective.

**Example: Data made accessible to organisations**

Organisation A creates a dataset by treating personal data in such a manner that identifiability risk is subject to particular controls. Organisation A intends to make the dataset available to Organisations B, C and D by storing the dataset in an environment secured with access controls.

In the hands of Organisation A, the data is likely to remain in scope of data protection law. For example, where it keeps the original personal data, or possesses the additional information that enables re-identification (even if held separately and subject to particular technical and organisational measures). Identifiability is reasonably likely (or already present) in both cases.

The status of the dataset may be different in the hands of Organisations B, C and D, depending on the circumstances. However, the key difference in this case is that Organisation A is not disclosing the dataset by transmission but it is making it available to those organisations.

Fundamentally, the nature of any assessment of identifiability risk is the same. The organisation making the data available may have additional considerations than in a one-to-one situation, and these should take place before the data is actually available to other parties.

For example, Organisation A might:

- consider which other organisations are likely to request access to the environment during the initial scoping and design stages. This may be relatively clear upfront, depending on the information itself and the likely purposes and nature of the organisations in question;
- ensure that the requesting organisations provide their own assessment of identifiability risk before giving access, so that the status of the information when they do access it is established;
- ensure that any other organisation seeking access to the data has a legitimate reason to do so and only accesses the minimum data needed to achieve their purpose;
- apply specific technical and organisational controls to particular organisations, if appropriate; and
- monitor access to the environment and periodically review the data being accessed as appropriate.

## When should we review our identifiability assessments?

At the early stage of any treatment of personal data, you need to think about whether the techniques you use today are likely to remain appropriate to manage identifiability risk in the future.

As noted above, the status of information can change over time. Technological developments may increase identifiability risk, moving it from sufficiently remote to reasonably likely. Information that is effectively anonymised today may become personal data in the future. Essentially, the information may ‘move’ along the spectrum of identifiability, so that data protection law starts to apply to it.

Ensuring you have the ability to consider these issues will also help you implement data protection by design effectively.

So, when you are considering and applying particular techniques, you should make realistic assessments relative to the circumstances of the case.

You should periodically review the decisions you take and the assessments that underpin them. You should do this at appropriate intervals. The timing and frequency of your review ultimately depends on the specifics of the information you anonymise, as well as the circumstances both of its disclosure and its use afterwards.

In some cases your review may need to be more continuous in nature. In others, it may be appropriate to review at particular points with longer intervals. There may also be particular events that lead to a review. For example, if a particular technological development means that a technique you originally used is no longer effective.

As noted above, you do not need to take account of every hypothetical or theoretical risk of identifiability. The key is whether identifiability is 'reasonably likely' given the circumstances. You need to make sure that, as technology changes, you update your original assessment to reflect the impact that change may have on your decision-making.

There are many sources of information available about current and future technologies, as well as existing and foreseeable threats. You therefore need to carry out periodic reviews of:

- your initial assessment;
- the technologies and techniques you use to render personal data into anonymous information;
- the state of technological development, and the steps you will take to account for it; and
- your overall policy for releasing data in light of the above.

You should undertake your reviews as appropriate to the circumstances of your processing and the likelihood of identifiability risk changing.

## Deciding when and how to release data

The considerations in this section of the guidance will help you ensure your assessment of identifiability risk is appropriate for the type of disclosure you undertake.

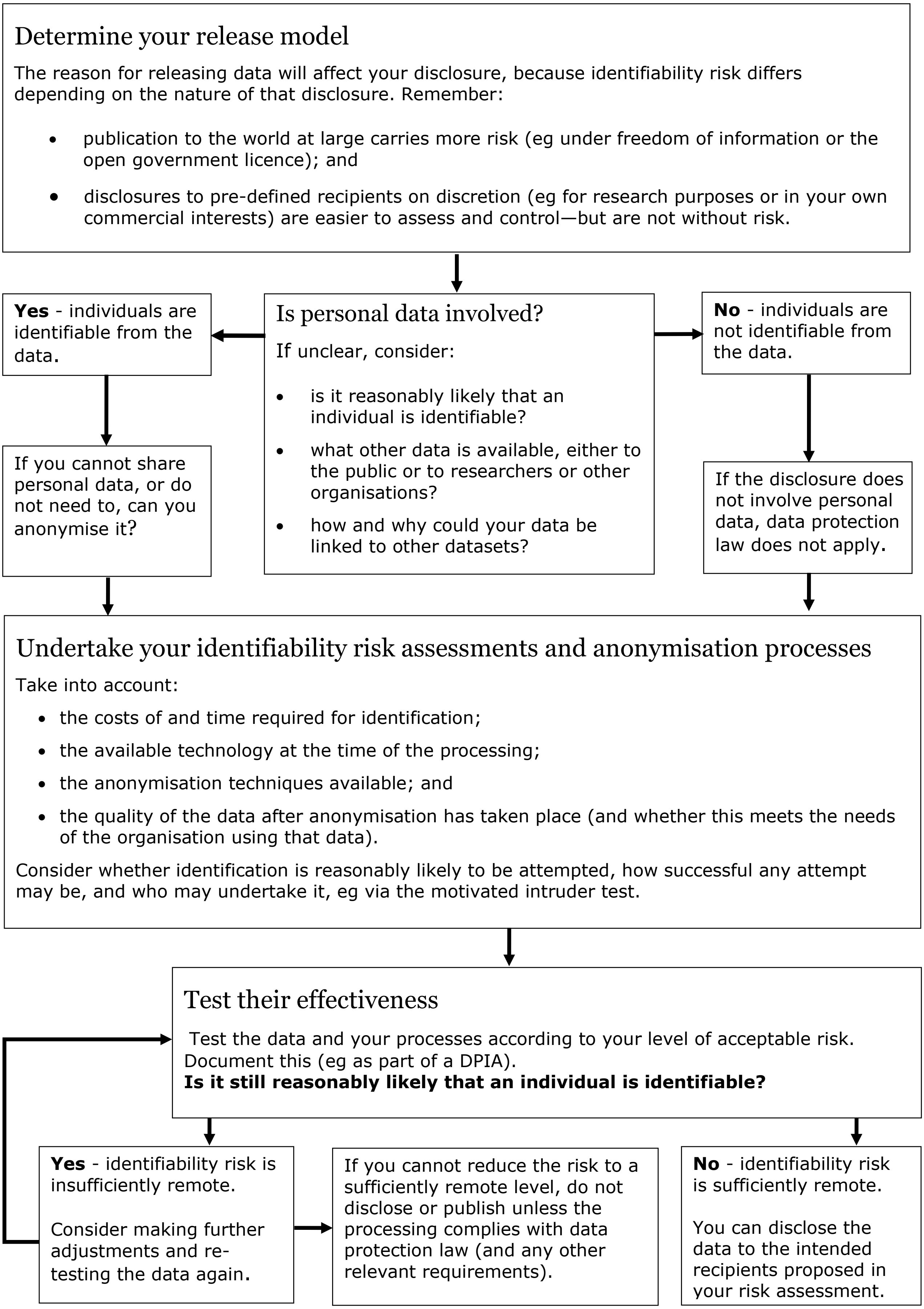
In summary, you should:

- determine your release model;
- conduct an initial assessment to assess whether the information includes personal data;
- establish whether you can anonymise that data;
- test the effectiveness your of anonymisation techniques, eg by assessing whether individuals are still identifiable;
- make further adjustments as appropriate; and
- document the above, including the decision you make about the disclosure.

Figure 2 below represents a way that you can implement this process.

**Further reading – ICO guidance**

Later sections of this guidance will also cover additional factors relating to accountability and governance in the context of anonymisation.



## Determine your release model

The reason for releasing data will affect your disclosure, because identifiability risk differs depending on the nature of that disclosure. Remember:

- publication to the world at large carries more risk (eg under freedom of information or the open government licence); and
- disclosures to pre-defined recipients on discretion (eg for research purposes or in your own commercial interests) are easier to assess and control—but are not without risk.

## Is personal data involved?

If unclear, consider:

- is it reasonably likely that an individual is identifiable?
- what other data is available, either to the public or to researchers or other organisations?
- how and why could your data be linked to other datasets?

**Yes** - individuals are identifiable from the data.

If you cannot share personal data, or do not need to, can you anonymise it?

**No** - individuals are not identifiable from the data.

If the disclosure does not involve personal data, data protection law does not apply.

## Undertake your identifiability risk assessments and anonymisation processes

Take into account:

- the costs of and time required for identification;
- the available technology at the time of the processing;
- the anonymisation techniques available; and
- the quality of the data after anonymisation has taken place (and whether this meets the needs of the organisation using that data).

Consider whether identification is reasonably likely to be attempted, how successful any attempt may be, and who may undertake it, eg via the motivated intruder test.

## Test their effectiveness

Test the data and your processes according to your level of acceptable risk. Document this (eg as part of a DPIA).
**Is it still reasonably likely that an individual is identifiable?**

**Yes** - identifiability risk is insufficiently remote.

Consider making further adjustments and re-testing the data again.

If you cannot reduce the risk to a sufficiently remote level, do not disclose or publish unless the processing complies with data protection law (and any other relevant requirements).

**No** - identifiability risk is sufficiently remote.

You can disclose the data to the intended recipients proposed in your risk assessment.