Data protection

Anonymisation: managing data protection risk code of practice

ico.
Information Commissioner's Office

Contents

Information Commissioner's foreword 4

1. About this code 6

2. Anonymisation and personal data 11

3. Ensuring anonymisation is effective 18

4. Do you need consent to produce or disclose anonymised data? 28

5. Personal data and spatial information 30

6. Withholding anonymised data 34

7. Different forms of disclosure 36

8. Governance 39

9. The Data Protection Act research exemption 44

Appendix 1 – Glossary 48

Appendix 2 – Some key anonymisation techniques 51

Appendix 3 – Further reading and sources of advice 54

Annex 1 – Converting personal data into anonymised data 57

Annex 2 – Anonymisation case-studies 66

Annex 3 – Practical examples of some anonymisation techniques 80

Information Commissioner's foreword

The UK is putting more and more data into the public domain. The government's open data agenda allows us to find out more than ever about the performance of public bodies. We can piece together a picture that gives us a far better understanding of how our society operates and how things could be improved. However, there is also a risk that we will be able to piece together a picture of individuals' private lives too. With ever increasing amounts of personal information in the public domain, it is important that organisations have a structured and methodical approach to assessing the risks. This code of practice is about managing that risk. My office has seen the risks both understated and overstated.

My office has been a strong supporter of the open data agenda, and has played its part in ensuring that all sorts of valuable data has been made available through the Freedom of Information Act 2000. One thing that has become clear, however, from my office's experience of dealing with information rights is that issues surrounding the release of information about individuals can be the hardest to deal with in practice. Finding out about the performance of a public authority, for example, inevitably involves finding out about the performance of its staff. We want openness, but we want privacy too. That is why the subject matter of this code of practice – anonymisation – is so important. If we assess the risks properly and deploy it in the right circumstances, anonymisation can allow us to make information derived from personal data available in a form that is rich and usable, whilst protecting individual data subjects.

The current Data Protection Directive, dating from 1995, says that the principles of data protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable. It also says that a code of practice can provide guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible. Yet, as far as I am aware, this is the first code of practice on anonymisation to be published by any European data protection authority. Issues surrounding anonymisation techniques and the status of anonymised data are becoming a key issue as discussion of the European Commission's proposal for a new data protection framework continues.

This code of practice is not a security engineering manual, nor does it cover every anonymisation technique. The Anonymisation Network will provide greater access to more detailed expertise and advice. But it does contain clear, practical advice and a straightforward explanation of some very tricky legal concepts. This code of practice will be of use to freedom of information and data protection practitioners, and to all those who are contributing to the creation of one of the world's most transparent and accountable economies.



Christopher Graham
Information Commissioner





About this code

Key points:

- Data protection law does not apply to data rendered anonymous in such a way that the data subject is no longer identifiable. Fewer legal restrictions apply to anonymised data.
- The anonymisation of personal data is possible and can help service society's information needs in a privacy-friendly way.
- The code will help all organisations that need to anonymise personal data, for whatever purpose.
- The code will help you to identify the issues you need to consider to ensure the anonymisation of personal data is effective.
- The code focuses on the legal tests required in the Data Protection Act.

The code explains the issues surrounding the anonymisation of personal data, and the disclosure of data once it has been anonymised. It explains the relevant legal concepts and tests in the Data Protection Act 1998 (DPA). The code provides good practice advice that will be relevant to all organisations that need to convert personal data into a form in which individuals are no longer identifiable. We use the term 'anonymised data' to refer to data that does not itself identify any individual and that is unlikely to allow any individual to be identified through its combination with other data.

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. Clearly, 100% anonymisation is the most desirable position, and in some cases this is possible, but it is not the test the DPA requires.

We use the term 're-identification' to describe the process of turning anonymised data back into personal data through the use of data matching or similar techniques. The code's annexes contain examples of various anonymisation and re-identification techniques and illustrations of how anonymised data can be used for various purposes. See Annex 1, which shows how a set of personal data can be converted into various forms of anonymised data.

We use the broad term 'anonymisation' to cover various techniques that can be used to convert personal data into anonymised data. We draw a distinction between anonymisation techniques used to produce aggregated information, for example, and those – such as pseudonymisation – that produce anonymised data but on an individual-level basis. The latter can present a greater privacy risk, but not necessarily an insurmountable one. We also draw a distinction between publication to the world at large and the disclosure on a more limited basis – for example to a particular research establishment with conditions attached. See case study 1: limited access to pharmaceutical data.

The code shows that the effective anonymisation of personal data is possible, desirable and can help society to make rich data resources available whilst protecting individuals' privacy. Anonymisation is of particular relevance at the moment, given the increased amount of information being made publicly available through Open Data initiatives and through individuals posting their own personal data online.

The code supports the Information Commissioner's view that the DPA should not prevent the anonymisation of personal data, given that anonymisation safeguards individuals' privacy and is a practical example of the 'privacy by design' principles that data protection law promotes. We hope that the code shows that, in some circumstances, anonymisation need not be an onerous process. In some cases really quite simple techniques can be very effective. See case study 2, using mobile phone data to study road traffic speeds and case study 3, which demonstrates a simple technique for anonymising data about passengers' travelling times.

Some information, particularly datasets containing sensitive personal data, will clearly present a need for caution, and the anonymisation issues may be complex for large datasets containing a wide range of personal data. It is in these complex scenarios in particular that organisations should consider whether they need specialist expertise and input.

This code was written for a general readership and only looks at the issue of anonymisation in the context of the DPA and Freedom of Information Act 2000 (FOIA). It does not go into all the other legal issues that could be relevant. We have tried to make the code as consistent as possible with other authoritative guidance. However, the Information Commissioner recognises that organisations may also need to follow their own detailed standards and procedures, tailored to the data they hold and its intended disclosure.

The code cannot describe every anonymisation technique that has been developed nor go into a great deal of technical detail. Additional information is available from the sources we have listed and will be developed through the Information Commissioner's Anonymisation Network. The Network will also host detailed case studies and illustrations of good practice. The network will be launched at the same time as this code of practice; details will be available on the ICO website.

Many important issues concerning anonymisation have arisen in the context of the FOIA and the Freedom of Information (Scotland) Act 2002 (FOISA). We are confident that this code will help public authorities in Scotland and the rest of the UK to deal with cases where personal data must be withheld, but anonymised data can be released. References to FOIA can be read across to include FOISA as well.

Who is this code of practice for?

Any organisation that needs or wants to turn personal data into anonymised data should use this code. This could be, for example, because the organisation:

- is required by law to publish anonymised data, such as some health service bodies;
- needs to deal with a request for information that contains third party personal data made under FOIA;
- wants to make itself more transparent and accountable to the public; or
- intends to further research or statistical purposes by making its anonymised data available to researchers.

Most of the good practice advice in the code will be applicable to public, private and third sector organisations, because the issues they face when anonymising personal data effectively are much the same. The majority of the code will apply to all instances of anonymisation regardless of its scale and context.

The code is not aimed at those seeking in-depth knowledge of security engineering or statistical methodology. However, the code will help experts in the field to understand the data protection framework their activities take place within.

How can the code help you?

Adopting the good practice recommendations in this code will give you a reasonable degree of confidence that your publication of anonymised data will not lead to an inappropriate disclosure of personal data – through 're-identification'.

The code will help you to identify the issues you need to consider when deciding how to anonymise personal data. It will also help you to assess any risk associated with producing – and particularly publishing – anonymised data.

These risks might include:

- information about someone's private life ending up in the public domain;
- an anonymised database being 'cracked' so that data about a number of individuals is compromised;

- individuals being caused loss, distress, embarrassment, or anxiety as a result of anonymised data being re-identified;
- reduced public trust if your organisation discloses anonymised data unsafely; and
- legal problems where insufficiently redacted qualitative data is disclosed, for example, under FOIA.

When the Information Commissioner investigates an issue relating to the anonymisation of personal data, he will take the good practice advice in this code into account. It will certainly stand an organisation in good stead if it can demonstrate that its approach to producing and disclosing anonymised data has taken account of the good practice recommendations set out in this code.

Specific benefits of this code include:

- minimising the risk of breaking the law and consequent enforcement action by the Information Commissioner's Office (ICO) or other regulators;
- promoting a better understanding of a difficult area of the law, particularly the data protection – freedom of information interface;
- developing a better understanding of anonymisation techniques, of the suitability of their use in particular situations and of their relative strengths and weaknesses;
- instilling greater confidence when dealing with UK-wide 'transparency agenda' imperatives for the publication of information – or with legal duties to publish;
- improving decision making when handling freedom of information requests involving personal data;
- developing greater public trust through ensuring that legally required safeguards are in place and are being complied with;
- reducing reputational risk caused by the inappropriate or insecure publication or disclosure of personal data; and
- reducing questions, complaints and disputes about your publication or disclosure of information derived from personal data.

Wider benefits of this code include:

- the furtherance of statistical and other research that relies on the availability of information derived from personal data;
- transparency as a result of organisations being able to make information derived from personal data available;
- the confidence to publish anonymised data in rich, re-usable formats;
- the economic benefits that the availability of rich data resources can bring;

- public confidence that data is being used for the public good whilst privacy is being protected; and
- better public authority accountability through the availability of data about service outcomes and performance.

The code's status

The Information Commissioner has issued this code under section 51 of the DPA in pursuance of his duty to promote good practice. The DPA says good practice includes, but is not limited to, compliance with the requirements of the DPA. This code was also published with Recital 26 and Article 27 of the European Data Protection Directive (95/46/EC) in mind. These provisions make it clear that the principles of data protection do not apply to anonymised data and open the way for a code of practice on anonymisation.

This code gives advice on good practice, but compliance with our recommendations is not mandatory where they go beyond the strict requirements of the DPA. The code itself does not have the force of law, as it is the DPA that places legally enforceable obligations on organisations.

Organisations may find alternative ways of meeting the DPA's requirements and of adopting good practice. However, if they do nothing then they risk breaking the law. The ICO cannot take enforcement action over a failure to adopt good practice or to act on the recommendations set out in this code unless this in itself constitutes a breach of the DPA.

We have tried to distinguish our good practice recommendations from the legal requirements of the DPA. However, there is inevitably an overlap because, although the DPA sets out the bare legal requirements, it provides no guidance on the practical measures that could be taken to comply with them. This code helps to plug that gap.

2 Anonymisation and personal data

Key points:

- Understanding anonymisation means understanding what personal data is.
- To protect privacy it is better to use or disclose anonymised data than personal data.
- It is possible to disclose anonymised data without breaching the Data Protection Act.

What is personal data?

The Data Protection Act 1998 (DPA) is concerned with 'personal data'. It says that 'personal data' means:

data which relate to a living individual who can be identified—

(a) from those data,

or

(b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual.

Personal data has to be about a living person, meaning that the DPA does not apply to mortality or other records about the deceased, although such data could still be protected by confidentiality or other legal rules.

What is not personal data?

From the definition above, it follows that information or a combination of information, that does not relate to and identify an individual, is not personal data. Clearly, effective anonymisation depends on a sound understanding of what constitutes personal data. See the Information Commissioner's Office (ICO)'s technical guidance on 'Determining what is personal data'.

Anonymisation in European data protection law

The most explicit reference to anonymisation in European data protection law is in Recital 26 of the European Data Protection Directive (95/46/EC) which:

- makes it clear that the principles of data protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable;
- recognises that a code of conduct, such as this one, can be a useful means of guidance as to how personal data can be rendered anonymous; and
- is particularly important because it indicates clearly that the anonymisation of personal data is to be considered possible and that it can be used to provide important privacy safeguards for individuals.

Anonymisation also supports data protection law's general data minimisation approach. Neither the Directive nor the DPA provide any technical advice on anonymisation techniques – which is why this code of practice should be particularly useful.

Note that the UK's DPA is framed in terms of identification or the likelihood of identification. The Data Protection Directive refers to 'likely reasonably'. In some cases the UK courts have used the 'likely reasonably' test. However, the practical problems that arise are much the same whether the test is of 'likelihood' of identification or 'reasonable likelihood' of it.

What are the benefits of anonymisation?

The DPA requires all organisations that process personal data to protect it from inappropriate use or disclosure. However, the same organisations may want, or be required, to publish information derived from the personal data they hold. For example, health service organisations are required to protect the identities of individual patients but may also be required to publish statistics about patient outcomes. Anonymisation helps organisations to comply with their data protection obligations whilst enabling them to make information available to the public.

Any organisation processing personal data has to comply with the data protection principles. The principles regulate the disclosure of personal data, and in some circumstances can prevent this. This means that, in general, it is easier to disclose anonymised data than personal data as fewer legal restrictions will apply. It is also easier to use anonymised data in new and different ways because the DPA's purpose-limitation rules do not apply to it. See case studies 8 and 9 for examples of how useful anonymised data can be.

Is anonymisation always necessary?

The primary reason for undertaking anonymisation is to protect individuals' privacy when making available the data resources that activities such as research and planning rely on. It is legitimate to use personal data for certain purposes, for example where the intention is to inform decisions about particular individuals, or to provide services to them. Much medical research involves access to patients' personal data and is carried out on the basis of patient consent and involvement. However, where the use of personal data is not necessary, then the objective should generally be to use anonymised data instead.

In some cases there will be no alternative to using personal data for research and certain other purposes. This might be the case for example where there is a need to contact individuals to ask them about the treatment they have received or the service they have subscribed to. The ICO recognises the special utility of personal data and that it is not always necessary or possible to use anonymised data instead of personal data.

Is anonymisation always possible?

The Information Commissioner recognises that some collections of personal data do not lend themselves well to anonymisation – eg voluminous collections of paper records held in a variety of formats. Although the sensitivity of data will generally decrease with the passage of time, the inappropriate release of records many decades old, eg criminal records, could still have a severely detrimental effect on an individual. That is why the security of data that cannot be anonymised is paramount. It is worth noting that the DPA's section 33 exemption – described later in this code - allows personal data held for research purposes to be retained indefinitely, provided certain conditions are met.

Disclosing anonymised data

There is clear legal authority for the view that where an organisation converts personal data into an anonymised form and discloses it, this will not amount to a disclosure of personal data. This is the case even though the organisation disclosing the data still holds the other data that would allow re-identification to take place. This means that the DPA no longer applies to the disclosed data, therefore:

- there is an obvious incentive for organisations that want to publish data to do so in an anonymised form;
- it provides an incentive for researchers and others to use anonymised data as an alternative to personal data wherever this is possible; and
- individuals' identities are protected.

A significant case relating to the anonymisation of personal data

R (on the application of the Department of Health) v Information Commissioner [2011] EWHC 1430 (Admin).

This case concerned the disclosure of medical statistics and whether they had been anonymised effectively.

In February 2005, the ProLife Alliance made a request under the Freedom of Information Act 2000 (FOIA) to the Department of Health for detailed statistical information about abortions carried in the year 2003. The ProLife Alliance specifically sought information about abortions carried out under 'Ground E' of the Abortion Act 1987, providing the same level of detail as set out in statistics provided by the Office of National Statistics (ONS) up until 2002.

Between 1968 and 2002 the ONS had published detailed information about Ground E abortions, being abortions carried out as there was a substantial risk that if the child were born it would suffer such physical or mental abnormalities as to be seriously handicapped. The ONS statistics listed a number of different foetal abnormalities and provided the total number of abortions for each one, together with a figure for terminations of over 24 weeks gestation. Responsibility for publishing abortion statistics was given to the Department of Health in 2002, which, in relation to the statistics for Ground E abortions, chose to combine certain categories of abnormality and suppress figures where the figure was between zero and nine.

The Department of Health refused the ProLife Alliance's request for the 2003 abortion statistics, providing the pre-2002 level of detail, relying on a number of FOIA exemptions from disclosure, including the exemption in section 40 concerning personal data. Following a complaint to the Information Commissioner and an appeal to the Information Tribunal, the matter was heard in the High Court before Mr Justice Cranston. The key consideration was whether the detailed abortion statistics were personal data for the purposes of the DPA.

The court referred to the DPA definition of personal data and Recital 26 of the European Data Protection Directive 95/46/EC which, in part, provides that "the principles of protection should not apply to data rendered anonymous in such a way that the data subject is no longer identifiable". Consideration was also given to the Article 29 working party Opinion (4/2007) on the concept of personal data. The Opinion concluded that anonymous data, in the sense used when applying the Directive, could be defined as any information relating to a natural person, where the person could not be identified, whether by the data controller or by any other person,

taking into account all means likely reasonably to be used to identify that individual.

Mr Justice Cranston, following the reasoning of Lord Hope in the case of Common Services Agency v Scottish Information Commissioner [2008] UKHL 47, held that, the fact that the data controller has access to all the information from which the statistical information is derived, does not disable it from processing the data in such a way, consistently with Recital 26 of the Directive, that it becomes data from which a living individual can no longer be identified. If converting the underlying information into statistics can achieve this, the way will then be open for the data controller to disclose the information in statistical form because it will no longer be personal data. Mr Justice Cranston held that the disclosure by the Department of Health of the detailed abortion statistics would not amount to the disclosure of personal data. In converting the underlying information into statistics, the Department of Health had effectively anonymised the information so that, taking account of all the means likely reasonably to be used, anyone receiving the statistics would not be able to identify any of the individuals to whom the statistics related.

Disclosing personal data

The DPA does not prohibit the disclosure of personal data, but any disclosure has to be fair, lawful and in compliance with the other data protection principles. The age of the information and level of detail can be important factors, for example data showing where individuals lived or worked sixty years ago may have little sensitivity in many cases. There is no hard and fast rule here, but a good rule of thumb is to try to assess the effect – if any - that the disclosure would have on any individual concerned and what their attitude to the disclosure would be likely to be. This could be influenced by whether the data is about their private life or about more public matters, such as their working life. See the ICO's Guide to Data Protection for more information about the disclosure of personal data and about the 'conditions for processing' personal data.

Anonymisation within organisations

The DPA is mainly concerned with the disclosure of personal data outside the data controller's own boundaries. However, anonymisation can also be relevant to the safe use or sharing of data within organisations, particularly large ones with diverse functions. For example, a retailer might use anonymised data rather than customer purchase records for its stock planning purposes.

Personal data and identification

The definition of 'personal data' can be difficult to apply in practice for two main reasons:

- the concept of 'identify' – and therefore of 'anonymise' - is not straightforward because individuals can be identified in a number of different ways. This can include direct identification, where someone is explicitly identifiable from a single data source, such as a list including full names, and indirect identification, where two or more data sources need to be combined for identification to take place; and
- you may be satisfied that the data your organisation intends to release does not, in itself, identify anyone. However, in some cases you may not know whether other data is available that means that re-identification by a third party is likely to take place.

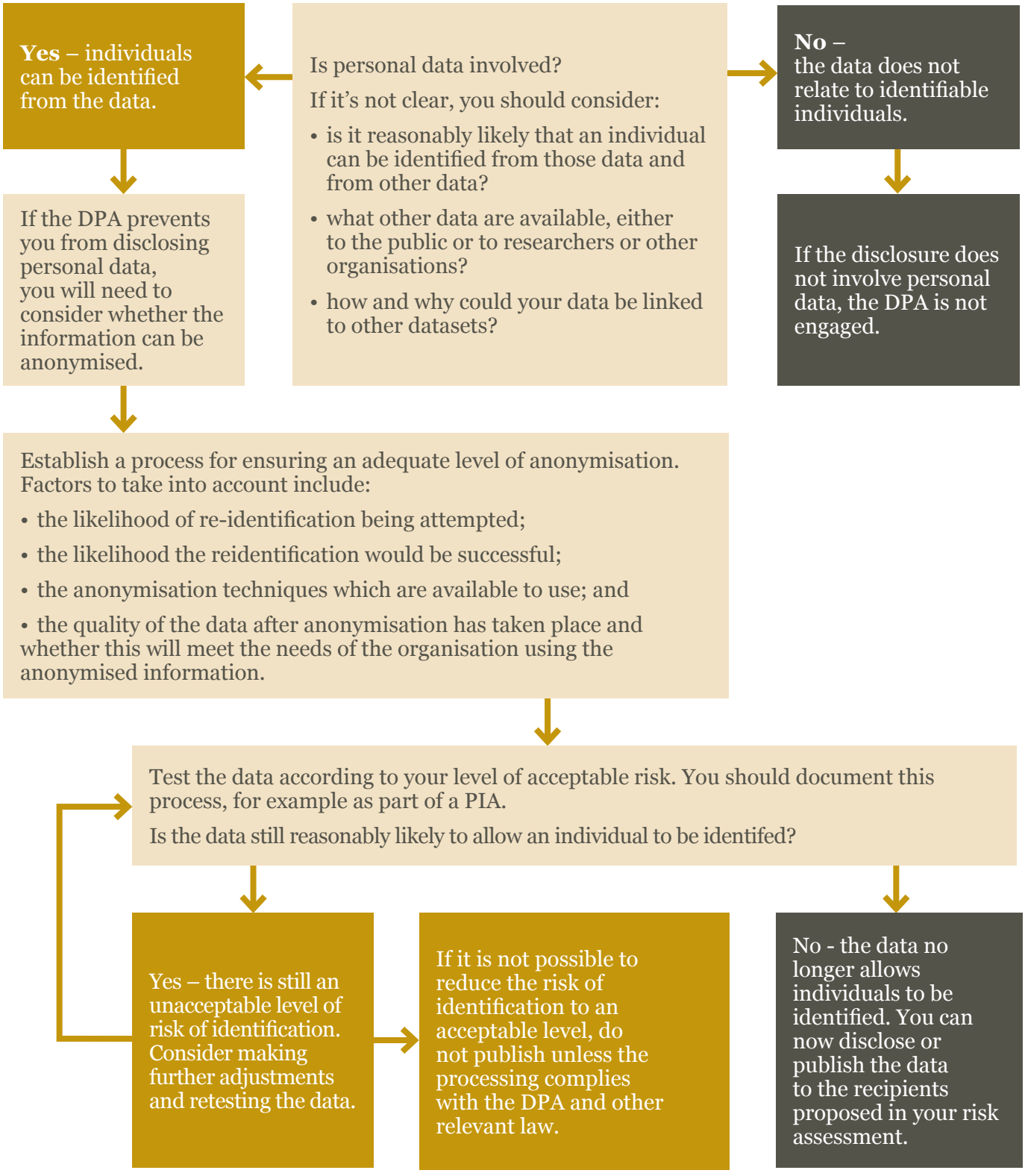
In reality it can be difficult to determine whether data has been anonymised or is still personal data. This can call for sensible judgement based on the circumstances of the case in hand. This code describes ways of assessing and mitigating the risks that may arise, particularly in terms of assessing whether other data is available that is likely to make re-identification likely. In some cases, it will be relatively easy to determine whether it is likely that a release of anonymised data will allow the identification of an individual. In other cases it will be much harder, but the decision still has to be made. See case study 4: publicly available information and anonymisation risk.

The DPA is not framed in terms of the *possibility* of an individual being identified. Its definition of personal data is based on the identification or likely identification of an individual. This means that, although it may not be possible to determine with absolute certainty that no individual will ever be identified as a result of the disclosure of anonymised data, this does not mean that personal data has been disclosed. The High Court in the Department of Health case above stated that the risk of identification must be greater than remote and reasonably likely for information to be classed as personal data under the DPA. See the summary of the R v Department of Health abortion statistics on page 14.

Deciding when and how to release anonymised data

The reason for releasing data will affect how you make the disclosure, because the risk and consequences of identification will differ:

- Publication under freedom of information or the open government licence is to the wider world, and carries more risk.
- Discretionary disclosures, such as those made for research purposes or in your own commercial interests, can be easier to control and assess but are not without risks.



3 Ensuring anonymisation is effective

Key points:

- It can be impossible to assess re-identification risk with absolute certainty.
- There will be many borderline cases where careful judgement will need to be used, based on the circumstances of the case.
- If you produce personal data through a re-identification process, you will take on your own data controller responsibilities.

What 'other' information is out there?

On the face of it, it can seem fairly easy to say whether a particular piece of information relates to and identifies an individual or not, and therefore whether it is personal data. Bank statements, for example, clearly identify individual account holders and contain information that relates to them.

However, the Data Protection Act 1998 (DPA) says that personal data means data which relate to a living individual who can be identified from those data, or from those data and *other information* which is in the possession of, or is likely to come into the possession of, the data controller. Determining what other information is 'out there', who it is available to and whether it is likely to be used in a re-identification process can clearly be extremely problematic.

The 'other information' needed to perform re-identification could be information available to certain organisations, to certain members of the public or that is available to everyone because it has been published on the internet, for example. Clearly the risk of combining information to produce personal data increases as data linkage techniques and computing power develop, and as more potentially 'match-able' information becomes publicly available.

It is worth stressing that the risk of re-identification through data linkage is essentially unpredictable because it can never be assessed with certainty what data is already available or what data may be released in the future. It is also generally unfeasible to see data return (ie recalling data or removing it from a website) as a safeguard given the difficulty, or impossibility, of securing the deletion or removal of data once it has been published. That is why it is so important to take great care, and to carry out as thorough a risk analysis as is possible, at the initial stage of producing and disclosing anonymised data.

There are two main ways for re-identification to come about.

- An intruder takes personal data it already has and searches an anonymised dataset for a match.
- An intruder takes a record from an anonymised dataset and seeks a match in publicly available information.

Generally the latter risk scenario is of greater concern for data custodians because of the confidentiality pledges that are often given to those appearing in an anonymised dataset. However, both risk scenarios are relevant and can carry with them different probabilities of re-identification. In either case though it can be difficult, even impossible, to assess risk with certainty.

Despite all the uncertainty, re-identification risk can certainly be mitigated by ensuring that only the anonymised data necessary for a particular purpose is released. The fact that data has been anonymised does not mean that data minimisation techniques are not still relevant.

Freedom of information and personal data

The DPA is primarily concerned with the risks associated with the identification of individuals by data controllers. However, section 40 of the Freedom of Information Act 2000 (FOIA) introduces a broader concept of risk because its test for deciding whether personal data can be disclosed is whether disclosure to *a member of the public* would breach the data protection principles. This means that public authorities have to assess whether releasing apparently anonymised data to *a member of the public* would breach the data protection principles. This is intended to ensure that public authorities take into account the additional information that a particular member of the public might have that could allow data to be combined to produce information that *relates to and identifies* a particular individual - and that is therefore personal data.

The test in FOIA can be particularly difficult to apply in practice because different members of the public may have different degrees of access to the 'other information' needed for re-identification to take place. However, a motivated intruder test can go some way towards addressing this problem.

It is good practice to try to look at identification 'in the round', ie all organisations disclosing anonymised data should assess whether any organisation or member of the public could identify any individual from the data being released – either in itself or in combination with other available information. The risk involved will vary according to the local data environment and particularly who has access to information. This means that anonymised data disclosed within a secure local environment, eg when disclosed to a particular research organisation, could remain anonymous even though if published, the likelihood of re-identification would mean that the anonymised data would become personal data. See case study 5 for an example of a FOIA decision notice relating to the disclosure of anonymised data.

What should I do in a borderline case?

There will clearly be borderline cases where, in reality, it will be difficult, or even impossible, to determine whether it is likely that re-identification will take place. The test in the DPA for determining whether information relating to a living individual is 'personal data' is based entirely on the identification or likely identification of the individual. The risk posed to individuals by disclosure, or the public benefit of this, are not factors that the DPA allows to be taken into account when determining whether or not information is personal data. In reality though, some types of data will be more attractive to a motivated intruder than others – and more consequential for individuals. In reality these factors should also inform an organisation's approach to disclosure.

Clearly the identification of an individual can have a range of consequences depending on the nature of the data, the context in which it is disclosed and who it is about. The Information Commissioner would certainly be more concerned about a disclosure of personal data that is detrimental to an individual, than about an inconsequential one. The Information Commissioner will take the effect or potential effect into account should a case of re-identification or inappropriate data disclosure come to his attention.

In borderline cases where the consequences of re-identification could be significant eg because they would leave an individual open to damage, distress or financial loss, organisations should:

- seek data subject consent for the disclosure of the data, explaining its possible consequences;
- adopt a more rigorous form of risk analysis and anonymisation.

In some scenarios, data should only be disclosed within a properly constituted closed community and with specific safeguards in place.

In some particularly high-risk situations, it may not even be possible to share within a closed community.

Even if a FOIA request is refused on section 40 (personal data) grounds, a more limited or possibly restricted form of disclosure might satisfy the requester. FOIA does not rule out this approach and it may help the requester if some anonymised data is released, rather than the request being turned down entirely on section 40 grounds. It may also reduce the risk, and expense, of an appeal.

It is worth noting that even if the 'likelihood' test points towards identification and the information is therefore personal data, you can still consider disclosure but will need to consider the other tests in the DPA, such as fairness. The DPA only prevents the disclosure of personal data under FOIA and more generally, where this would breach the data protection principles.

What is the risk of re-identification?

In some cases the risk of anonymised data being combined with other data to result in personal data being created will be high. An obvious example is where publicly available data – such as the Electoral Roll or data easily retrievable from a web-search – can be combined with the 'anonymised' data, allowing an individual to be identified. Note that 'identified' does not necessarily mean 'named'. It can be enough to be able to establish a reliable connection between particular data and a known individual.

However, in some circumstances it can be difficult to establish the risk of re-identification, particularly where complex statistical methods might be used to match various pieces of anonymised data. This can be a particular vulnerability where pseudonymised data sets are concerned, because even though pseudonymised data does not identify an individual, in the hands of those who do not have access to the 'key', the possibility of linking several anonymised datasets to the same individual can be a precursor to identification. This does not mean though, that effective anonymisation through pseudonymisation becomes impossible. The Information Commissioner recognises that some forms of research, for example longitudinal studies, can only take place where different pieces of data can be linked reliably to the same individual. The DPA does not prevent this provided that:

a) identification does not take place, or

b) if identification does take place, this does not constitute a breach of the data protection principles.

The principles would be breached if individuals were assured that only anonymised data would be published but in fact their personal data was disclosed.

Data controllers must be aware of the risk of re-identification and that this risk can change over time, eg powerful data analysis techniques that were once rare are now common-place. However, if anonymisation is carried out effectively in the present this is likely to protect personal data from future re-identification.

A realistic assessment of the risk of re-identification occurring in the future should be made, meaning that organisations should not assume that data that is anonymous now will necessarily become re-identifiable in the future. However, organisations should carry out a periodic review of their policy on the release of data and of the techniques used to anonymise it, based on current and foreseeable future threats. There are certainly examples though of where a complacent approach to anonymisation, and insufficiently rigorous risk analysis, has led to the substantial disclosure of personal data. This was the case where 'anonymised' internet search results were released without proper consideration of the risk of individuals identifying each other from the search terms used.

The risk of one anonymised dataset being matched with another to produce personal data can be reduced by using sampling techniques,

so that only parts of databases rather than whole ones are released – making direct linkage more difficult.

Anonymising qualitative data

Much of the anonymised data being created, used and disclosed is derived from administrative datasets that are essentially statistical in nature. However, the techniques used to anonymise quantitative data are not generally applicable when seeking to anonymise qualitative data, such as the minutes of meetings, interview transcripts or video footage. Different techniques are needed to do this. Obvious methods include:

- redacting individuals' names from documents;
- blurring video footage to disguise faces;
- electronically disguising or re-recording audio material; and
- changing the details in a report (precise place names, precise dates etc.)

Inevitably, the anonymisation of qualitative material can be time-consuming. It does not lend itself to bulk processing and can require careful human judgement based on the data in question. The sections of this code that deal with assessing re-identification risk will be helpful here. The UK Data Archive also provides guidance on the anonymisation of qualitative data. See case study 6 for an example of anonymised qualitative data.

The 'motivated intruder' test

Neither the DPA nor the FOIA provide any practical assistance in terms of helping organisations to determine whether:

a) the anonymised data they release is likely to result in the re-identification of an individual; or

b) whether anyone would have the motivation to carry out re-identification.

However a useful test – and one used by the Information Commissioner and the Tribunal that hears DPA and FOIA appeals – involves considering whether an 'intruder' would be able to achieve re-identification *if* motivated to attempt this.

The 'motivated intruder' is taken to be a person who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been derived. This test is meant to assess whether the motivated intruder would be successful.

The approach assumes that the 'motivated intruder' is reasonably competent, has access to resources such as the internet, libraries, and all public documents, and would employ investigative techniques such as making enquiries of people who may have

additional knowledge of the identity of the data subject or advertising for anyone with information to come forward. The 'motivated intruder' is not assumed to have any specialist knowledge such as computer hacking skills, or to have access to specialist equipment or to resort to criminality such as burglary, to gain access to data that is kept securely.

Clearly, some sorts of data will be more attractive to a 'motivated intruder' than others. Obvious sources of attraction to an intruder might include:

- finding out personal data about someone else, for nefarious personal reasons or financial gain;
- the possibility of causing mischief by embarrassing others;
- revealing newsworthy information about public figures;
- political or activistic purposes, eg as part of a campaign against a particular organisation or person; or
- curiosity, eg a local person's desire to find out who has been involved in an incident shown on a crime map.

However, this does not mean that data that is, on the face of it, 'ordinary', 'innocuous' or without value can be released without a thorough assessment of the threat of re-identification.

In some cases there may be a high level of risk to individuals should re-identification occur. One example might be health data, where, although there may be no obvious motivation for trying to identify the individual that a particular patient 'episode' relates to, the degree of embarrassment or anxiety that re-identification could cause could be very high. Therefore, the anonymisation techniques used to protect data should reflect this. In reality though, data with the potential to have a high impact on an individual is most likely to attract a 'motivated intruder'.

The 'motivated intruder' test is useful because it sets the bar for the risk of identification higher than considering whether a 'relatively inexpert' member of the public can achieve re-identification, but lower than considering whether someone with access to a great deal of specialist expertise, analytical power or prior knowledge could do so. It is therefore good practice to adopt a 'motivated intruder' test as part of a risk assessment. Carrying out a motivated intruder test in practice might include:

- carrying out a web search to discover whether a combination of date of birth and postcode data can be used to reveal a particular individual's identity;
- searching the archives of national or local newspaper to see whether it is possible to associate a victim's name with crime map data;
- using social networking to see if it is possible to link anonymised data to a user's profile; or

- using the electoral register and local library resources to try to link anonymised data to someone's identity.

It is good practice to periodically re-assess the risk of re-identification through motivated intrusion, bearing in mind that as computing power and the public availability of data increases, so will the re-identification risk. Where re-identification results in the processing of personal data, the organisation doing the processing will take on its own data protection responsibilities. See case study 4 for an example of how publicly available information can aid re-identification.

Motivated intruder risk: some issues to consider

- What is the risk of jigsaw attack, ie piecing different bits of information together to create a more complete picture of someone? Does the information have the characteristics needed to facilitate data linkage - eg is the same code number used to refer to the same individual in different datasets?
- What other 'linkable' information is available publicly or easily?
- What technical measures might be used to achieve re-identification?
- How much weight should be given to individuals' personal knowledge?
- If a penetration test has been carried out, what re-identification vulnerabilities did it reveal?

Obvious sources of information include:

- Libraries
- Local council offices
- Church records
- General Registry Office
- Genealogy websites
- Social media; internet searches
- Local and national press archives
- Anonymised data releases by other organisations, particularly public authorities

Prior knowledge and re-identification

Re-identification problems can arise where one individual or group of individuals already knows a great deal about another individual, for example a family member, colleague, doctor, teacher or other professional. These individuals may be able to determine that

anonymised data relates to a particular individual, even though an 'ordinary' member of the public or an organisation would not be able to do this. Examples of this include:

- a doctor knowing that an anonymised case study in a medical journal relates to a patient she is treating;
- one family member knowing that an indicator on a crime map relates to an assault another family member was involved in; and
- an employee working out that an absence statistic relates to a colleague who he knows has been on long-term sick leave.

The risk of re-identification posed by making anonymised data available to those with particular personal knowledge cannot be ruled out, particularly where someone might learn something 'sensitive' about another individual – if only by having an existing suspicion confirmed. However, the privacy risk posed could, in reality, be low where one individual would already require access to so much information about the other individual for re-identification to take place. Therefore a relevant factor is whether the other individual will learn anything new. An example of this might be an individual's genetic code. This would identify an individual uniquely, but only to someone who already has access to both the code and the identity of the individual it belongs to. The situation is similar where an individual might recognise that anonymised data relates to him or her, allowing self-identification to take place. See case study 7, which shows how prior knowledge can be a factor in re-identification.

It is important not to make assumptions about family relationships when considering prior knowledge and what individuals may already know. The most obvious example is certain medical information teenagers may not share with their parents or other family members.

It is good practice when releasing anonymised data to try to assess:

- the likelihood of individuals having and using the prior knowledge necessary to facilitate re-identification. It is accepted that this will be difficult to conduct on a record by record basis for large datasets or collections of information. It will often be acceptable to make a more general assessment of the risk of prior knowledge leading to identification, for at least some individuals recorded in the information and then make a global decision about the information; the chances that those who might be able to re-identify are likely to seek out or come across the relevant data; and
- what the consequences of re-identification are likely to be, if any, for the data subject concerned. Of course this can be difficult to assess in practice and a member of the public's sensitivity may be different from yours. For example, the disclosure of the address of a person on a witness protection scheme could be far more consequential than would usually be the case.

It is reasonable to conclude that professionals (such as doctors) with prior knowledge are not to be likely to be motivated intruders, if it is clear their profession imposes confidentiality rules and requires ethical conduct.

Information, established fact and knowledge

When considering re-identification risk, it is useful to draw a distinction between recorded information, established fact and personal knowledge:

- Established fact might be that Mr B Stevens lives at 46 Sandwich Avenue, Stevenham. This could have been established by looking at an up-to-date copy of the electoral register.
- Personal knowledge might be that I know Mr B Stevens is currently in hospital, because my neighbour – Mr Stevens' wife – told me so.

The starting point for assessing re-identification risk should be recorded information and established fact. It is easier to establish that particular recorded information is available, than to establish that an individual – or group of individuals - has the knowledge necessary to allow re-identification. However, there is no doubt that non-recorded personal knowledge, in combination with anonymised data, can lead to identification. It can be harder though to substantiate or argue convincingly. There must be a plausible and reasonable basis for non-recorded personal knowledge to be considered to present a significant re-identification risk.

Identification and the educated guess

Data protection law is concerned with information that identifies an individual. This implies a degree of certainty that information is about one person and not another. Identification involves more than making an educated guess that information is about someone; the guess could be wrong. The possibility of making an educated guess about an individual's identity may present a privacy risk but not a data protection one because no personal data has been disclosed to the guesser. Even where a guess based on anonymised data turns out to be correct, this does not mean that a disclosure of personal data has taken place. However, the consequences of releasing the anonymised data may be such that a cautious approach should be adopted, even where the disclosure would not amount to a disclosure of personal data. Therefore it may be necessary to consider whether the data should be withheld for some other reason, as discussed later in this code.

This is clearly a difficult area of the law and in approaching questions of disclosure it can be helpful to look primarily at the possible impact on individuals and then to move on to the more technical issue of whether or not there is likely to be a disclosure of personal data subject to the DPA.

Information about groups of people

In some circumstances the release of anonymised data can present a privacy risk even if it does not constitute personal data and cannot be converted back into personal data. This might be the case where the anonymised data points to a number of individuals,

eg the occupants of a group of households or those living within a particular postcode area. Information that enables a group of people to be identified, but not any particular individual within the group is not personal data. Conversely, information that does enable particular individuals within a group – or all the members of a group – to be identified will be personal data in respect of all the individuals who can be identified. There is no doubt that releasing information about groups of people can give rise to privacy and other risks. An obvious example would be where released information indicated that someone living in a small geographical area had committed a serious crime. Even though that individual is not identifiable, there might be a health and safety risk to all those in the area if reprisals were likely.

Even if public authorities cannot rely on the 'personal data' exemption in FOIA to prevent the release of information like this, they may be able to rely on other exemptions, bearing in mind that the public interest may favour disclosure where an exemption is not absolute. Organisations that are not public authorities should also adopt an approach of balancing the risk that disclosure may pose to an individual or group of individuals against the benefit that might result from disclosure.

What if you create personal data from anonymised data?

Initiatives such as open data, and the publication on the internet of information released under FOIA, mean that it is easier than ever to 'harvest' and analyse large amounts of data. This will include anonymised data derived from personal data and personal data itself.

This means that the opportunity may arise for a 'motivated intruder' individual or organisation to combine, analyse and match publicly available data to create personal data anew or to link additional data to existing personal data, eg to find out more about a person.

If an organisation collects or creates personal data then it will take on its own data protection responsibilities in respect of the data. This could require it to inform the individuals concerned that data about them is being processed. This could clearly present reputational or legal problems, particularly where individuals would not expect your organisation to have personal data about them, or may find this objectionable.

The Information Commissioner will generally take the view that where an organisation collects personal data through a re-identification process without individuals' knowledge or consent, it will be obtaining personal data unlawfully and could be subject to enforcement action.

The Information Commissioner is confident that adopting the techniques and procedures recommended in this code will guard against re-identification. However, in some cases re-identification may be a possibility. Where there is evidence of re-identification taking place, with a risk of harm to individuals, the Information Commissioner will be likely to take regulatory action, including the imposition of a civil monetary penalty of up to £500,000.

4 Do you need consent to produce or disclose anonymised data?

Key points

- Consent is generally not needed to legitimise an anonymisation process.
- Even if consent can be obtained it is usually 'safer' to use or disclose anonymised data.
- The Information Commissioner's Office recognises that obtaining consent can be very onerous or even impossible.

Do I need consent?

The Data Protection Act 1998 (DPA) provides various 'conditions' for legitimising the processing of personal data, including its anonymisation. Consent is just one condition, and the DPA usually provides alternatives. The DPA only gives the individual a right to prevent the processing of their personal data where this would be likely to cause unwarranted damage or distress. In the Information Commissioner's Office (ICO)'s view, it follows therefore that provided there is no likelihood of anonymisation causing unwarranted damage or distress – as will be the case if it is done effectively – then there will be no need to obtain consent as a means of legitimising the processing.

When is consent viable?

The publication of personal data based on an individual's properly informed consent will not breach the data protection principles. Certainly, organisations that involve individuals and obtain their consent for the creation and disclosure of their personal data can stay inside the law and can build up an open and positive relationship with the individuals whose personal data they are processing. This could be the case, for example, where individuals agree to take part in a valuable but potentially quite intrusive longitudinal health study.

Obtaining consent for the anonymisation of personal data can be logistically very onerous, eg where large numbers of personal records are involved. It could even be impossible – eg where the personal data is old and there is no reliable means of contacting individual data subjects. This is often the case with historical archives which may contain the personal data of individuals who are still living.

What happens if consent is withdrawn?

However, there can be problems in an approach based on consent, particularly where this involves the publication of personal data. If an individual can give consent, the individual can withdraw it – and may want to do so because of a change in their personal circumstances, for example. Even if the withdrawal of consent stops the original data controller from further processing the personal data, in reality, it may be impossible to remove the data from the public domain. The withdrawal of consent may have little or no effect. It is therefore 'safer' to publish anonymised data than personal data, even where consent could be obtained for the disclosure of personal data itself.

Consent and obtaining personal data

It is important to consider how the personal data you wish to anonymise was obtained originally. If, for example, the data was collected as part of a survey and individuals were told that it would be used for research purposes then clearly there will be no barrier to using the data for that purpose. In some cases individuals may have been given an assurance that personal data about them would only be used for a particular purpose, eg to despatch the goods they have ordered. Assurances of this nature should be respected, but very specific purpose limitation of this type is rare.

A more common scenario is for an organisation to have a collection of personal data obtained for a particular purpose or set of purposes, eg to administer individuals' library services. In cases like this individuals may never have been notified as to whether their data will or will not be anonymised for use in research purposes, for example. Organisations should address this in their privacy policies or by other means.

Anonymising personal data obtained under an enactment

In some cases it is a legal requirement that personal data is provided to an organisation. For example, employers are generally required to file tax returns about their employees with HMRC. However, even if individuals have no choice over the provision of their personal data, this does not mean that they have the right to stop the organisation anonymising it – provided the processing of their personal data in order to anonymise is not likely to cause unwarranted damage or distress. Of course the processing of personal data must also comply with the data protection principles, meaning it must be 'fair', for example. Note that some official bodies such as central government departments, may be subject to specific legal constraints on the use of the personal data – and other data assets - they hold.

5 Personal data and spatial information

Key points

There is no simple rule for handling spatial information – such as postcodes, GPS data or map references - under the Data Protection Act 1998 (DPA). In some circumstances this will constitute personal data, eg where information about a place or property is, in effect, also information about the individual associated with it. In other cases it will not be personal data.

The context of the related information and other variables, such as the number of households covered by a postcode, is key. It is clear, though, that the more complete a postcode - or the more precise a piece of geographical information - the more possible it becomes to analyse it or combine it with other information, resulting in personal data being disclosed. However, where spatial information is being published for a legitimate purpose, the objective should be to achieve the maximum level of detail that can be balanced with the protection of individuals' privacy. A Privacy Impact Assessment (PIA) should be carried out to help you to do this.

The approach you should take to spatial information will also be guided by the size of the dataset you have; in some cases you may need to consider the position on a case by case basis. For example, this may be possible where a Freedom of Information Act 2000 (FOIA) request is for specific information linked to a postcode. In one decision, the Information Commissioner decided that burglary information linked to a particular postcode was not personal data – see Decision Notice FS50161581. In other cases you will have to take more global decisions about the status of different types of postcode or other spatial information.

In some cases it may be necessary to process spatial information to remove or 'blur' certain elements, to reduce the risk of identification. For example, in England, when anonymising postcodes the following average characteristics of postcodes should be considered:

- full postcode = approx 15 households (although some postcodes only relate to a single property)
- postcode minus the last digit = approx 120/200 households
- postal sector = 4 outbound digits + 1 inbound gives approx 2,600 households
- postal district = 4 outbound digits approx 8,600 households

- postal area = 2 outbound digits approx 194,000 households

  Source: Centre for Advanced Spatial Analysis: UCL

('Outbound' is the first part of the postcode, 'inbound' the second part; for example with the postcode SV3 5AF, the outbound digits are SV3 and the inbound digits are 5AF.)

An alternative approach – and one that may result in more useful data and avoid the problems of inaccuracy and misinterpretation that the use of partial postcodes can create – is to use 'replacement' postcodes for real ones. This may allow researchers to retain the granularity and accuracy of data whilst minimising re-identification risk when publishing data on a postcode basis. However, this approach will not be feasible when publishing data for public use, given that individuals will want to find out information referenced according to real postcode areas.

With information relating to a particular geographical area, there can be a distinction between a "statistical comfort zone" that eliminates almost all risk of identification, and other forms of information that pose a risk of an individual being identified. Small numbers in small geographical areas present increased risk, but this does not mean that small numbers should always be removed automatically. For example, always removing numbers relating to five or 10 individuals or fewer may be a reasonable rule of thumb for minimising the risk of identification in a proactive disclosure scenario, but in the context of a specific freedom of information request a different approach may be possible, based on an application of the tests in the DPA.

It is important that organisations consider the different geographical units used in other anonymised disclosures. One organisation may disclose data linked to postcode, others by ward level. As far as they can, organisations that disclose anonymised spatial datasets regularly should work together to assess the risk of jigsaw identification through overlapping geographical units. The Office for National Statistics website contains a useful guide to geographical units.

Smart phones and GPS

Mobile devices such smart phones and GPS systems generate significant amounts of detailed spatial information. It depends on how the systems operate as to whether the spatial information identifies and relates to an individual and is personal data. Often, many organisations are involved in delivering the different layers of services on these devices, so identification issues become more complex.

Organisations should consider how other unique identifiers (eg IP addresses) and other identifying information (eg names, addresses) are linked to the spatial information. In some circumstances organisations who offer services related to smartphones or GPS will be processing personal data. The answer may be different, depending on what other information the organisation using the spatial information has access to.

Individuals using smart phones will often be given an option to allow their device to reveal to its location - to the device itself or a particular application 'app'. It should be clear to the individual how this information is used by the device or the app. Privacy policies should clearly set out whether spatial information is processed as personal data and when it is only used in an anonymised form.

The concept of 'degrading' or 'fading' personal data is useful for organisations using spatial information. An organisation may need to process spatial information as personal data initially to enable a transaction to work, but once this has finished the need for precise information may have passed. Subsequently details could be replaced incrementally by more general information. For example, information about a user's exact GPS coordinates could be swapped for a street name, then a ward and then just a city.

The Information Commissioner's Office (ICO) has produced specific guidance on crime mapping. The following principles – developed from that guidance - are useful when considering the disclosure of geographical-based datasets to the public. We anticipate that the UK Anonymisation Network (UKAN, www.ukanon.net) will also contribute to the debate about the release of spatial data.

You can reduce the risk to privacy when publishing spatial information by:

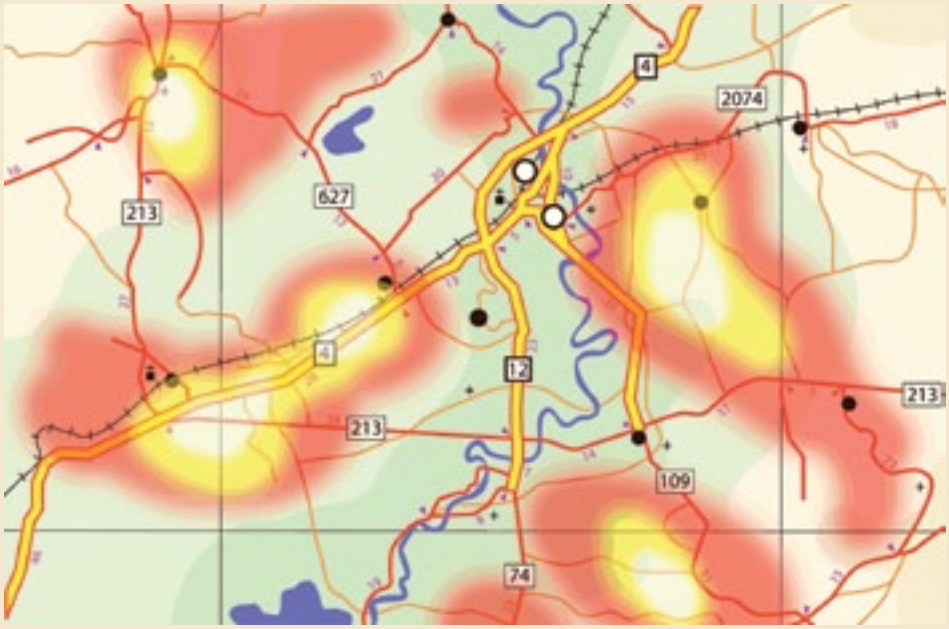
- increasing a mapping area to cover more properties or occupants;
- reducing the frequency or timeliness of publication, so that it covers more events, is harder to identify a recent case, or does not reveal additional data such as time or date of the event. Publishing data very frequently or in real-time poses a greater privacy risk;
- removing the final 'octet' on IP addresses to degrade the location data they contain;
- using formats, such as heat maps, that provide an overview without allowing the inference of detailed information about a particular place or person; and
- avoiding the publication of spatial information on a household level. This could constitute the processing of personal data because it is quite easy to link a property to its occupant or occupants – using the publicly available Electoral Register, for example.

Where there are no risks, or they are minimal, geographical information should provide as much information as possible, to enable the public to understand issues such as crime in their area. This can enable communities to engage with agencies such as the police and bring about enhanced accountability.

The risks that can emerge from the disclosure of geographical information are still emerging. As more data becomes available, as data-linkage tools develop and as computing power increases, the impact of disclosures of anonymised geographical datasets should be kept under review.

A heat map approach to crime mapping

The advantages of this are that there is no clear link, actual or suggested, between levels and types of crime and particular locations. This avoids misleading representation, for example where all the crimes occurring in a particular area are mapped to a smaller area or specific place. Heat mapping also makes it much more difficult for the general public to establish a link between a particular crime and a particular individual.



6 Withholding anonymised data

Key points

- The fact that data is not personal data does not mean you can always disclose it.
- The Data Protection Act's definition of personal data cannot be extended to cover situations where the data does not identify any individual.
- Public authorities need to consider their compliance with human rights law.

Organisations may want to disclose data that is not personal data. Clearly the Data Protection Act 1998 (DPA) will not prevent this. However, there may still be reasons for withholding data that is not personal data. Disclosing certain data could still present a risk to individuals, even if they cannot be identified from it. For example, a risk may arise where an educated guess leads to the *misidentification* of an individual. For example, available data plus individual knowledge might lead someone to believe that an innocent person was responsible for a particular crime.

The definition of personal data should not be extended to cover scenarios where no information that relates to an identifiable individual is involved. In the case of public authorities receiving a Freedom of Information Act (FOIA) request, another exemption may allow the information to be withheld. For example, FOIA's section 38 'health and safety' exemption could be relevant here.

The same considerations will apply when considering disclosure under the Freedom of Information (Scotland) Act 2002.

Human rights

It goes beyond the scope of this code to provide exhaustive guidance on the Human Rights Act (HRA). However, public authorities and private sector organisations - insofar as they carry out functions of a public nature – must comply with the HRA. Organisations subject to the HRA must not act in away that would be incompatible with rights under the European Convention on Human Rights. This includes Article 8 – the right to respect for private and family life. However, this is not an absolute right: public authorities are permitted to interfere with it where it is necessary, lawful and proportionate to do so.

The Article 8 right will often overlap with the protection provided for by the DPA; if a disclosure is compliant with the DPA it is likely to be compliant with the HRA. However, the Article 8 right is not limited to situations involving the processing of personal data. This means that some disclosures of information that do not engage the DPA could still engage the broader provision in the HRA. For example, information about a large family group might not be personal data but its disclosure may well breach the privacy rights of the family. It is advisable to seek specialist advice if you believe a disclosure has novel or potentially contentious Article 8 implications.

Other statutory prohibitions

Other statutory prohibitions may apply to the disclosure of information, with different tests and considerations to the DPA. For example, there are relatively strict limitations on the purposes for which certain government departments are allowed to produce and disclose even anonymised data. A breach of a statutory prohibition would engage FOIA's section 44 exemption.

Statistical confidentiality

Producers of Official and National Statistics must observe the Code of Practice for Official Statistics, and the related National Statistician's guidance on confidentiality.



7 Different forms of disclosure

Key points

- Different forms of anonymised data can pose different re-identification risks.
- Publication is more risky than limited access.
- Limited access allows the disclosure of 'richer' data.
- Limited access relies on robust governance arrangements.

Different types of anonymised data, different risks

A problem faced by those using anonymised data is that on the one hand they want data that is rich and usable enough for their purposes. On the other, they want to ensure that re-identification does not occur. This means that different disclosure options may need to be considered.

Different types of anonymised data have different vulnerabilities and pose different levels of re-identification risk. At one end of the spectrum, pseudonymised or de-identified data may be very valuable to researchers because of its individual-level granularity and because pseudonymised records from different sources can be relatively easy to match. However, this also means that there is a relatively high re-identification risk. At the other end of the spectrum, aggregated data is relatively low-risk, depending on granularity, sample sizes and so forth. This data may be relatively 'safe' because re-identification risk is relatively low. However, this data may not have the level of detail needed to support the data linkage or individual-level analysis that some forms of research depend on.

Given the very different types of anonymised data that can be derived from personal data, it is important for data controllers to consider their disclosure options carefully, ie does the data need to be published or would limited access be appropriate? In general, the more detailed, linkable and individual-level the anonymised data is, the stronger the argument for ensuring only limited access to it. This might be the case where it is necessary to use individual, record-level anonymised data to track particular individuals' movement through the education, employment and criminal justice systems.

The more aggregated and non-linkable the anonymised data is, the more possible it is to publish it. This might be the case for statistics showing the percentage of children in a wide geographical area who have achieved particularly high educational attainment, for example.

Publication versus limited access

It is important to draw a distinction between the publication of anonymised data to the world at large and limited access. Clearly the open data agenda relies on the public availability of data, and information released in response to a freedom of information request cannot be restricted to a particular person or group. However, much research, systems testing and planning, for example, takes place by releasing data within a closed community, ie where a finite number of researchers or institutions have access to the data and where its further disclosure is prohibited, eg by a contract. The advantage of this is that re-identification and other risks are more controllable, and potentially more data can be disclosed without having to deal with the problems that publication can cause. It is therefore important to draw a clear distinction between:

- publication to the world at large, eg under the Freedom of Information Act 2000 or open data. Here – in reality - there is no restriction on the further disclosure or use of the data and no guarantee that it will be kept secure; and
- limited access, eg within a closed community of researchers. Here it is possible to restrict the further disclosure or use of the data and its security can be guaranteed.

Limited access is particularly appropriate for the handling of anonymised data derived from sensitive source material or where there is a significant risk of re-identification.

There can still be risks associated with limited access disclosure - but these can be mitigated where data is disclosed within a closed community working to established rules. Data minimisation rules will also remain relevant.

It could be appropriate that data anonymised from a collection of personal data is published, whilst a record-level version of the data is released in a limited way under an end-user agreement.

Limited access safeguards

The organisation responsible for the initial disclosure of the data on a limited access basis must put robust safeguards in place before the data can be made available to others. These should include:

- purpose limitation, ie the data can only be used by the recipient for an agreed purpose or set of purposes;
- training of recipients' staff with access to data, especially on security and data minimisation principles;
- personnel background checks for those getting access to data;
- controls over the ability to bring other data into the environment, allowing the risk of re-identification by linkage or association to be managed;

- limitation of the use of the data to a particular project or projects;
- restriction on the disclosure of the data;
- prohibition on any attempt at re-identification and measures for the destruction of any accidentally re-identified personal data;
- arrangements for technical and organisational security, eg staff confidentiality agreements;
- encryption and key management to restrict access to data;
- limiting the copying of, or the number of copies of the data;
- arrangements for the destruction or return of the data on completion of the project; and
- penalties, such as contractual ones that can be imposed on the recipients if they breach the conditions placed on them.

It should be noted a pre-defined list of risk mitigations cannot be exhaustive. Data controllers must conduct their own risk assessment, eg using their organisation's normal data security risk assessment processes. Co-ordination between the organisations involved in a project should help to identify other security measures that may need to be included.

Publication under licence

Once data has been published under a licence - such as the Open Government Licence - it may be impossible to protect it from further use or disclosure or to keep it secure. However, the Open Government Licence does make it clear that while anonymised data falls within the scope of the licence, users and re-users are not permitted to use the data in a way that enables re-identification to take place. However, this may be difficult or impossible to enforce.

8 Governance

Key points

- Organisations anonymising personal data need an effective and comprehensive governance structure.
- The ICO will ask about your governance if we receive a complaint or carry out an audit.
- There needs to be senior-level oversight of your governance arrangements.

If your organisation is involved in the anonymisation and disclosure of data, it is good practice to have an effective and comprehensive governance structure in place that will address the practical issues surrounding the production and disclosure of anonymised data.

Having an effective governance structure in place will help you if the Information Commissioner's Office (ICO) receives a complaint about your processing of personal data, including its anonymisation, or if we carry out an audit. Enforcement action – including the imposition of monetary penalties - is less likely where an organisation can demonstrate that it has made a serious effort to comply with the Data Protection Act (DPA) and had genuine reason to believe that the data it disclosed did not contain personal data or present a re-identification risk.

A governance structure should cover the following areas.

- Responsibility for authorising and overseeing the anonymisation process. This should be someone of sufficient seniority and with the technical and legal understanding to manage the process. A 'Senior Information Risk Owner' (SIRO) approach can be particularly useful. The role of the SIRO is to take responsibility for key decisions and to inform an organisation's general corporate approach to anonymisation. A SIRO should be able to coordinate a corporate approach to anonymisation, drawing on relevant expertise from within and outside an organisation. The SIRO should be able to help its organisation decide on suitable forms of disclosure, ie publication or limited access.
- Staff training: staff should have a clear understanding of anonymisation techniques, any risks involved and the means of mitigating these. In particular, individual staff members should understand their specific roles in ensuring anonymisation is being done safely.
- Procedures for identifying cases where anonymisation may be problematic or difficult to achieve in practice: These could be cases

where it is difficult to assess re-identification risk or where the risk to individuals could be significant. It is good practice to have procedures in place to identify these difficult cases and to document how a decision was made as to how, or whether, to anonymise the personal data and how, or whether, to disclose it.

- Knowledge management regarding any new guidance or case law that clarifies the legal framework surrounding anonymisation. Knowledge management should also extend to new techniques that are available to organisations anonymising data and to intruders seeking to identify individuals within a dataset. Participating in the ICO's Anonymisation Network will be a good way to develop understanding, to assess risk and to share expertise.

- A joined up approach with other organisations in their sector or those doing similar work. Organisations should seek to share information about planned disclosures with other organisations, to assess risks of jigsaw identification. For example it would be helpful for public authority A to know that public authority B is also planning an anonymised disclosure at the same time, one on health and one on welfare, both using similar geographical units. They can then assess the risks collectively and agree mitigation for both datasets.

- Privacy impact assessment (PIA): This is an effective method of assessing privacy risks in a structured way. A PIA could contain elements intended to test the effectiveness of an anonymisation technique, helping you to assess re-identification risk to devise mitigation measures. Many organisations involved in the creation or disclosure of anonymised data will find the ICO's PIA handbook a useful way to structure and document their decision-making process. The approach in the handbook can easily be read across to cover many anonymisation scenarios. The Information Commissioner recommends that organisations should normally publish their PIA report to show the public how they have approached the risk assessment process.
  Read the ICO PIA handbook

- Transparency. As anonymised data has no direct effect on any individual, there can be a tendency not to tell individuals about it, or even to be secretive. It may not be necessary, and in many cases will be impossible, to contact individual data subjects. However, your organisation's privacy policy – which should be clear and easily accessible to the public - should explain your organisation's approach to anonymisation as clearly as possible and any consequences of this. In particular your organisation should:

  - explain why you anonymise individuals' personal data and describe in general terms the techniques that will be used to do this;

  - make it clear whether individuals have a choice over the anonymisation of their personal data, and if so how to exercise this – including the provision of relevant contact details. (Note though that the DPA does not give individuals a general right to prevent the processing of personal data about them);

- say what safeguards are in place to minimise the risk that may be associated with the production of anonymised data. In particular, you should explain whether the anonymised data will be made publicly available or only disclosed to a limited number of recipients;
- be open with the public about any risks of the anonymisation you are carrying out – and the possible consequences of this. You should give them the opportunity to submit queries or comments about this; and
- describe publicly the reasoning process regarding the publication of anonymised data, explaining how you did the 'weighing-up', what factors you took or did not take into account and why, how you looked at identification 'in the round'. This mode of transparency should improve trust as well as lead to improvements in the decision process itself through exposure to public scrutiny and comment.

  Whilst it is good practice to be as transparent as possible, you should not disclose data that would make re-identification more likely. However, excessive secrecy is likely to generate public distrust and suspicion.

  Organisations should also consider whether they can publish any PIA reports on anonymisation, removing certain information if needed or publishing a summary report.

- Review of the consequences of your anonymisation programme, particularly through the analysis of any feedback you receive about it. Review should be an on-going activity and 're-identification testing' techniques should be used to assess re-identification risk and to mitigate this. It is important to analyse and deal with any complaints or queries you receive from members of the public who believe that their privacy has been infringed.
- Disaster recovery: your governance procedures should also address what you will do if re-identification does take place and individuals' privacy is compromised. This could involve telling individuals there has been a breach and helping them to take any necessary remedial action. A re-identification incident may lead to the cessation of the anonymisation process or to its modification, eg by using more rigorous anonymisation techniques or disclosure controls.

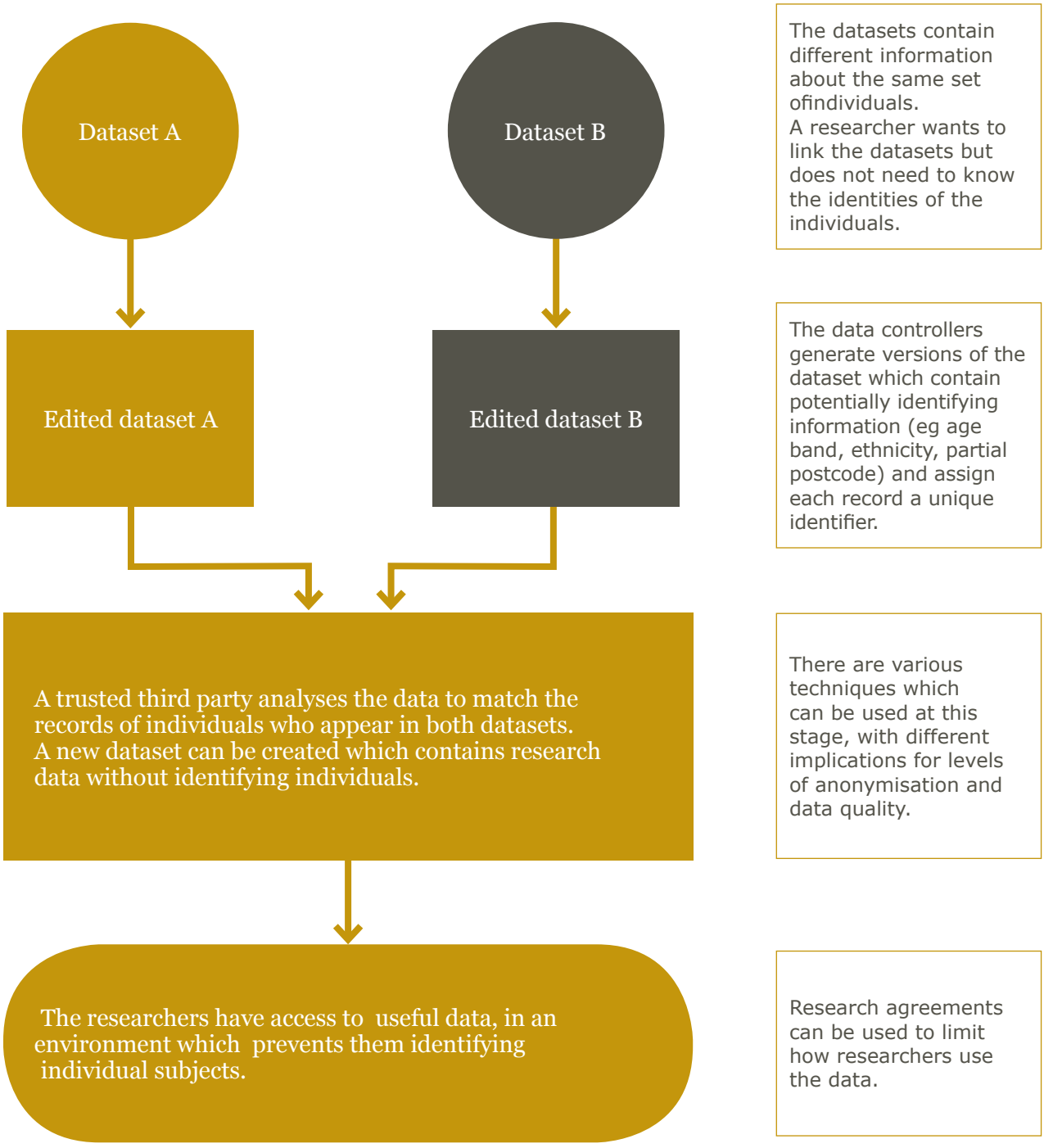
Trusted third parties

A trusted third party (TTP) arrangement can be particularly effective where a number of organisations each want to anonymise the personal data they hold for use in a collaborative project. This model is being used increasingly to facilitate the large scale research using data collected by a number of organisations.

Typically, the TTP will operate a data repository to which the various participating organisations will disclose their personal data.

Using a trusted third party to anonymise data

A trusted third party is an organisation which can be used to convert personal data into an anonymised form. This is particularly useful in the context of research, as it allows researchers to use anonymised data in situations where using raw personal data is not necessary or appropriate. Trusted third parties can be used to link datasets from separate organisations, and then create anonymised records for researchers.



The personal data can then be anonymised in 'safe', high security conditions and to an agreed specification – allowing the subsequent linkage of anonymised individual-level data, for example.

The great advantage of a TTP arrangement is that it allows social science research to take place – for example using anonymised data derived from health and criminal justice records – without the organisations involved ever having access to each others' personal data. Security, anonymisation and anti-re-identification measures taken by the TTP should be covered on agreement.

Re-identification testing

It is good practice to use re-identification testing – a type of 'penetration' or 'pen' testing - to detect and deal with re-identification vulnerabilities. This involves attempting to re-identify individuals from an anonymised data set or data sets.

There can be advantages in using a third party organisation to carry out the testing, as it may be aware of data resources, techniques or types of vulnerability that you have overlooked or are not aware of.

The first stage of a re-identification testing process should be to take stock of the anonymised data that your organisation has published or intends to publish. The next stage should be to try to determine what other data - personal data or not - is available that could be linked to the anonymised data to result in re-identification. As we have explained elsewhere, this can be difficult to do in practice because, in reality, it may be difficult or impossible to determine what other information particular individuals or organisations will have access to. However, you should certainly check whether other publicly available information is available – or is easily accessible through a web-search, for example – that could allow re-identification to take place. The 'motivated intruder' test described earlier can form a useful component of a pen-test.

A penetration test should meet the following criteria:

- the test should attempt to identify particular individuals and one or more private attributes relating to those individuals.
- the test may employ any method which is reasonably likely to be used by an intruder.
- the test may use any lawfully obtainable data source which is reasonably likely to be used to identify particular individuals in the datasets.

Assessing re-identification risk becomes more complex where statistical data is involved, because the various statistical data sets may be publicly available which, if matched in a particular way, could result in re-identification. There could also be a risk of re-identification using the data within a particular dataset itself. Pen-testing for this type of vulnerability can require a great deal of specialist knowledge and cannot be described fully in this code of practice. The UK Anonymisation Network (UKAN) will explore this topic further.

9 The Data Protection Act research exemption

Key points

- The Data Protection Act's research exemption contains limited but useful features for researchers.
- Researchers processing personal data still have to comply with most of the DPA, including its principles.

The Data Protection Act (DPA) contains an exemption that relates to personal data processed only for research purposes. It is relevant to this code of practice because much anonymised data is used for research purposes.

What does the DPA say?

Section 33 of the DPA says this:

Research, history and statistics.

(1) In this section— "research purposes" includes statistical or historical purposes; "the relevant conditions", in relation to any processing of personal data, means the conditions—

(a) that the data are not processed to support measures or decisions with respect to particular individuals, and

(b) that the data are not processed in such a way that substantial damage or substantial distress is, or is likely to be, caused to any data subject.

(2) For the purposes of the second data protection principle, the further processing of personal data only for research purposes in compliance with the relevant conditions is not to be regarded as incompatible with the purposes for which they were obtained.

(3) Personal data which are processed only for research purposes in compliance with the relevant conditions may, notwithstanding the fifth data protection principle, be kept indefinitely.

(4) Personal data which are processed only for research purposes are exempt from section 7 if—

(a) they are processed in compliance with the relevant conditions, and

(b) the results of the research or any resulting statistics are not made available in a form which identifies data subjects or any of them.

(5) For the purposes of subsections (2) to (4) personal data are not to be treated as processed otherwise than for research purposes merely because the data are disclosed—

(a) to any person, for research purposes only,

(b) to the data subject or a person acting on his behalf,

(c) at the request, or with the consent, of the data subject or a person acting on his behalf, or

(d) in circumstances in which the person making the disclosure has reasonable grounds for believing that the disclosure falls within paragraph (a), (b) or (c).

What is 'research'?

The DPA does not define 'research'. Therefore the Information Commissioner will use an ordinary meaning of 'research' when determining whether personal data is being processed for research purposes: research is a systematic investigation intended to establish facts, acquire new knowledge and reach new conclusions.

The DPA makes it clear that 'research purposes' include statistical or historical research, but other forms of research, for example market, social, commercial or opinion research, could benefit from the exemption.

What sort of data is section 33 relevant to?

The exemption is clearly of most relevance where personal data – rather than anonymised data – is being used for research. The exemption is as applicable to sensitive personal data, eg data about someone's health being processed for medical research – as it is to 'ordinary' personal data. It provides important - though limited - assistance to those seeking to use personal data for research purposes. As explained elsewhere in this code, it is not always possible to use anonymised data for research purposes. Therefore researchers should be aware of the useful features that this exemption contains and the protection for individuals that it provides.

The exemption can apply to data collected primarily for research purposes and to cases where research is a secondary purpose.

However, the part of the exemption that deals with incompatibility is clearly of most relevance where research is a secondary purpose.

Section 33 safeguards

For the exemption to apply, certain conditions must be satisfied:

- the data must not be processed to support measures or decisions with respect to particular individuals.
- the data must not be processed in such a way that substantial damage or substantial distress is, or is likely to be, caused to any data subject.

Where anonymisation is carried out effectively, neither the production nor the publication of the anonymised data will have any effect on any particular individual. Provided that this is the case, the research exemption's conditions will have been satisfied.

Incompatibility, retention and subject access

Provided the data is only processed for research purposes, and the conditions are satisfied, then:

- the data may be processed for research purposes without falling foul of the DPA's prohibition on processing data for an 'incompatible' purpose. This puts it beyond doubt that personal data obtained for one purpose can also be used for research purposes;
- the data may be retained indefinitely. This is important in contexts such as historical research or longitudinal studies because the data protection principles usually require that personal data is not kept for longer than is necessary. Note that the data protection principles do not apply to anonymised data; and
- the data will be exempt from the right of subject access – provided the data is not published in a form which identifies any individual or individuals. This means that organisations can avoid the administrative issues associated with dealing with individuals' requests. It is good practice though to grant individuals access to personal data held for research purposes even if the exemption does apply.

Clearly the research exemption provides important benefits for researchers and important safeguards for individuals. However, it is good practice to plan for the publication of anonymised data as early in the data life cycle as is practicable. This will help to minimise, or will negate, the risk to individuals. It also means that researchers will not need to be concerned with the parts of the DPA from which section 33 does not provide exemption, eg the requirement to process personal data fairly and lawfully. See case study 10 for an example of anonymisation being used in a longitudinal study.

The disclosure of research data

The section 33 exemption can still be relied on even if research outputs are published in a form which identifies individuals, but the exemption from providing subject access will be lost. However, depending on the circumstances, the publication of personal data for research purposes could still breach other provisions of the DPA.

There is a particular incentive to anonymise sensitive personal data, eg data about someone's health or criminal convictions. This is because this type of personal data is subject to relatively stringent data protection restrictions. In particular, it could be difficult to find an alternative to seeking the data subject's consent as a means of

legitimising the processing of sensitive data about their health. (In some cases organisation may, as a matter of policy, decide to always obtain data subjects' consent for the anonymisation of personal data about them, but the DPA provides alternatives to this.) This is why anonymisation should occur at the earliest opportunity – ideally by the data controller anonymising the personal data prior to disclosing or using it for research purposes.

The DPA does not necessarily prohibit the disclosure of research data in a form which identifies individuals and the benefit of the section 33 exemption will not necessarily be lost if this happens. However, even if a researcher needs personal data to carry out research, it is arguably a breach of the DPA to publish or disclose data for research purposes in a form which identifies individuals where there is an alternative to this. Remember that an organisation that receives personal data from a researcher will take on its own data protection responsibilities as the data controller for that data. This could mean informing the individuals concerned that your organisation has obtained personal data about them.

If an individual consents to the use or disclosure of personal data about them for research purposes then there will be no need to rely on the DPA's research exemption. However, it can be impossible for organisations or individuals to exercise control over personal data once it has been published. An obvious problem might be where an individual who once consented to the use or disclosure of their personal data decides to revoke consent, eg because of a change in their personal circumstances. Therefore it is generally better to use and disclose anonymised data rather than personal data for research and other purposes - even where consent could be obtained. (It is rare for research outputs to be published in the form of personal data and consent for this would not normally be sought for this type of disclosure.)



Appendix 1 – Glossary

Aggregated data: Statistical data about several individuals that has been combined to show general trends or values without identifying individuals within the data.

Anonymisation: The process of rendering data into a form which does not identify individuals and where identification is not likely to take place.

Anonymised data: Data in a form that does not identify individuals and where identification through its combination with other data is not likely to take place.

Data controller: A person who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any personal data are, or are to be, processed.

Data linkage: A technique that involves bringing together and analysing data from a variety of sources, typically data that relates to the same individual.

Data processor: An organisation that processes personal data on behalf of a data controller.

Data subject: An individual who is the subject of personal data.

Disclosure: The act of making data available to one or more third parties.

Disclosure Control: A technique used to control the risk of individuals being identified from statistical data – typical methods include removing or disguising data relating to individuals with unusual sets of attributes.

Limited access: Releasing data within a closed community – i.e. where a finite number of researchers or institutions have access to the data and where its further disclosure is prohibited.

Longitudinal study: A study that involves linking data about the same individual over a period of time, eg to study an individual's health episodes.

Open Data: The government's white paper defines Open Data as:

Data that meets the following criteria:

- **accessible** (ideally via the internet) at no more than the cost of reproduction, without limitations based on user identity or intent;

- in a **digital, machine readable** format for interoperation with other data; and
- **free of restriction on use or redistribution** in its licensing conditions.

Personal data: Data which relate to a living individual who can be identified—

(a) from those data, or

(b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller,

and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual.

Perturbation: The alteration of values within a data set to guard against data-linkage.

Pseudonymisation: The process of distinguishing individuals in a dataset by using a unique identifier which does not reveal their 'real world' identity.

Publishing: The act of making data publicly available.

Qualitative data: Data gathered and analysed in a non-numeric form, such as interview transcripts, field notes, video and audio recordings, still images and documents such as reports, meeting minutes, e-mails etc.

Re-identification: The process of analysing data or combining it with other data with the result that individuals become identifiable. Sometimes termed 'de-anonymisation'.

Sensitive personal data: Personal data consisting of information as to—

(a) the racial or ethnic origin of the data subject,

(b) his political opinions,

(c) his religious beliefs or other beliefs of a similar nature,

(d) whether he is a member of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992),

(e) his physical or mental health or condition,

(f) his sexual life,

(g) the commission or alleged commission by him of any offence, or

(h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.

Statistical data: Information which is held in the form of numerical data, nominal data (eg gender, ethnicity, region), ordinal data (age group, qualification level), interval data (month of birth) or ratio data (age in months).



Appendix 2 – Some key anonymisation techniques

Data masking

This involves stripping out obvious personal identifiers such as names from a piece of information, to create a data set in which no person identifiers are present.

Variants:

- **Partial data removal** – results in data where some personal identifiers, eg name and address have been removed but others such as dates of birth, remain.
- **Data quarantining** - The technique of only supplying data to a recipient who is unlikely or unable to have access to the other data needed to facilitate re-identification. It can involve disclosing unique personal identifiers – eg reference numbers – but not the 'key' needed to link these to particular individuals.

These are relatively high risk techniques because the anonymised data still exists in an individual-level form. Electoral roll data, for example, could be used to reintroduce names that have been removed to the dataset fairly easily. However, this type of data is also relatively 'rich' in terms of allowing an individual to be tracked as part of a longitudinal study for example.

Pseudonymisation

De-identifying data so that a coded reference or pseudonym is attached to a record to allow the data to be associated with a particular individual without the individual being identified.

Deterministic modification is a similar technique. 'Deterministic' here means that the same original value is always replaced by the same modified value. This means that if multiple data records are linked, in the sense that the same name (or address, or phone number, for example) occurs in all those records, the corresponding records in the modified data set will also be linked in the same way. This facilitates certain types of data analysis.

This is also a relatively high risk technique, with similar strengths and weaknesses to data masking.

Aggregation

Data is displayed as totals, so no data relating to or identifying any individual is shown. Small numbers in totals are often suppressed through 'blurring' or by being omitted altogether.

Variants:

- Cell suppression - if data is from a sample survey then it may be inappropriate to release tabular outputs with cells which contain small numbers of individuals, say below 30. This is because the sampling error on such cell estimates would typically be too large to make the estimates useful for statistical purposes. In this case, suppression of cells with small numbers for quality purposes acts in tandem with suppression for disclosure purposes.
- Inference Control – Some cell values (eg small ones such as 1-5) in statistical data can present a greater risk of re-identification. Depending on the circumstances, small numbers can either be suppressed, or the values manipulated (as in Barnardisation). If a large number of cells are affected, the level of aggregation could be changed. For example, the data could be linked to wider geographical areas or age-bands could be widened.
- Perturbation – such as Barnardisation - is a method of disclosure control for tables or counts. It involves randomly adding or subtracting 1 from certain cells in the table. This is a form of perturbation.
- Rounding – rounding a figure up or down to disguise precise statistics. For example if one table may have a cell with value of 10,000 for all people doing some activity up to the present date. However, the following month, the figure in that cell rises to 10,001. If an intruder compares the tables it would be easy to deduce a cell of 1. Rounding would prevent this.
- Sampling - in some cases, when very large numbers of records are available, it can be adequate for statistical purposes to release a sample of records, selected through some stated randomized procedure. By not releasing specific details of the sample, data holders can minimise the risk of re-identification.

- Synthetic data - mixing up the elements of a dataset – or creating new values based on the original data - so that all of the overall totals and values of the set are preserved but do not relate to any particular individual.
- Tabular reporting – a means of producing tabular (aggregated) data, which protects against re-identification.
- These are relatively low risk techniques because it will generally be difficult to find anything out about a particular individual by using aggregated data. This data cannot support individual-level research but can be sufficient to analyse social trends on a regional basis, for example.

Derived data items and banding

Derived data is a set of values that reflect the character of the source data, but which hide the exact original values. This is usually done by using banding techniques to produce coarser-grained descriptions of values than in the source dataset eg replacing dates of birth by ages or years, addresses by areas of residence or wards, using partial postcodes or rounding exact figures so they appear in a normalised form.

Again, this is a relatively low-risk technique because the banding techniques make data-matching more difficult or impossible. The resulting data can be relatively rich because it can facilitate individual-level research but presents relatively low re-identification risk.

Appendix 3 – Further reading and sources of advice

Administrative Data Liaison Service: useful advice and resources for researchers, including guidance on privacy protection techniques.

A Systematic Review of Re-Identification Attacks on Health Data

Avoiding the Jigsaw Effect: 'Experiences With Ministry of Justice Reoffending Data'. Work carried out by Kieron O'Hara et al at the University of Southampton.

Class based graph anonymisation for social network data

Clinical Practice Research Datalink: www.cprd.com – advice on the use of anonymised NHS data

Data Anonymization and Re-identification: Some Basics Of Data Privacy

Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy (Ann Cavoukian and Khaled El Emam).

DWP / ESRC generic security accreditation document relating to explicit personal data and data that has not been sufficiently anonymised to make it freely available to the public.

Economic and Social Data Service: www.esds.ac.uk – see in particular its data management guides.

Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modelling Performance.

Government Social Research codes and guidance: www.civilservice.gov.uk/networks/gsr/publications.

Government Statistical Service: authoritative advice for government bodies about the creation and publication of statistical data.

ICO seminar on privacy and data anonymisation.

ICO website for advice on 'determining what is personal data', 'crime mapping', privacy by design, privacy enhancing technolgies and other issues relevant to the anonymisation of personal data.

Independent Privacy and Transparency Review (Kieron O'Hara)

Inference Control in Statistical Databases: From Theory to Practice (Lecture Notes in Computer Science)

Introduction to Privacy-Preserving Data Publishing Concepts and Techniques

Office for National Statistics – www.ons.gov.uk. In particular see its 'Guidance and Methodology' section. Also see its Code of Practice for Official Statistics.

Patient data for health research: A discussion paper on anonymisation procedures for the use of patient data for health research.

Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference, PSD 2008, Istanbul, Turkey, September 24-26, 2008

Protecting Privacy Using k-Anonymity

Statistical Confidentiality (2011) by G. Duncan, M. Elliot and J Salazar

Statistical Disclosure Control (2012) by A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E.Schulte Nordhold, K. Spicer and P.P de Wolf.

The problem of 'personal data' in cloud computing. International Data Privacy Law paper on anonymisation in the cloud. http://idpl.oxfordjournals.org/content/1/4/211.full

UK Data Archive: Practical advice on the legal, ethical and practical aspects of using data to carry out research – see in particular the guidance on anonymisation and access control.

UK Statistics Authority: Code of Practice for Official Statistics at www.statisticsauthority.gov.uk

Examples and case studies

These examples and cases studies are divided into three sections and are intended to illustrate the good practice advice contained in the main body of the code. Many of the examples were provided to us by respondents to the consultation exercise that preceded the publication of this code of practice.

The first annex consists of a detailed description of how a set of personal data can be converted into various forms of anonymised data and used in various ways. It also illustrates the difference between publication and limited disclosure and explores re-identification risk.

The second annex consists of case studies showing how various anonymisation techniques can be used in practice.

Finally, the third annex consists of a set of practical examples of some anonymisation techniques drawn up for the Information Commissioner's Office by experts at the University of Southampton.



Annex 1– research data held by the University of Stevenham Research Centre (USRC)

This case study shows how a collection of personal data about a group of individuals can be turned into various types of anonymised data. It also explores re-identification risk and shows how anonymised data derived from the personal data of the same individuals can be matched without their identities being revealed.

USRC's Public Health Research Department is investigating the relationship between the period of receipt of Special Assistance Benefit and individuals' age and body mass index (BMI).

USRC has collected a large amount of data including the following extract:

1. Name, address, date of birth	2. Period on Special Assistance benefit	3. Body mass index	4. Research cohort reference no.
Mr B Stevens 46 Water St Stevenham 20-4-69	1y 2m	15	1A5
Mrs C Davids 48 Water St Stevenham 18-3-60	5y 3m	14	2B4
Mr D Michaels 50 Water St Stevenham 16-2-71	1y 7m	16	3C3
Mrs E Seniuk 52 Water St Stevenham 14-1-62	5y 2m	18	4D2
Mr F O'Reilly 54 Water St Stevenham 12-12-63	1y 8m	20	5E1

Figure 1: Personal data

In the hands of USRC all of this collection of information constitutes personal data because all the data items relate to identified individuals. If it is disclosed to a third party in this form, this will be a disclosure of personal data and will be subject to the data protection principles.

A redacted data-set

USRC receives a freedom of information request from a neighbouring research centre (NRC), doing similar research on the relationship between individuals' time on Special Assistance Benefit, age range and BMI. USRC decides to disclose the following redacted data-set:

1. Name, address, date of birth	2. Period of Special Assistance Benefit.	3. Body mass index	5. Age range	6. Research cohort reference no.
	< 2 years	15	40-45	1A5
	> 5 years	14	50-55	2B4
	< 2 years	16	40-45	3C3
	> 5 years	18	45-50	4D2
	< 2 years	20	45-50	5E1

Figure 2: Information redacted from personal data

In creating the extract, USRC will be processing personal data. However, this will not breach the data protection principles as the purpose of the redaction process is to protect the individual research subjects' privacy and the processing itself has no direct effect on any individual.

The redacted data-set is still personal data in the hands of USRC because it still holds the full version of the original research data. This could act as a 'key' that would allow the extracted data to be linked back to personal identifiers – in this case individuals' names and addresses.

The extract is only be personal data in the hands of USRC because only USRC holds the 'key' needed to make the link back to the personal identifiers it holds. NRC cannot do this because there is no information in the extract itself that could allow the linkage to be made. This shows that at the point at which USRC discloses the extract, it ceases to be personal data – even though it is still personal data in the hands of USRC as long as it holds 'the other information' necessary to enable identification. (If USRC deletes

the full version of the research data – which only USRC holds - the extract will cease to be personal data in its, or anyone else's hands.)

Note that the 'research cohort reference number' – the unique identifier that USRC allocates to each individual involved in the research – cannot act as a 'key' for any organisation because only USRC has the complete set of information. However, other researchers may be able to use the number to individuate particular individuals without identifying them, or even to carry out longitudinal studies using USRC's annual data releases.

Aggregated or statistical data

USRC receives a further request for information, this time from a government agency that is planning service provision in the health service. It wants to know how many individuals that have claimed Special Assistance Benefit for less than two years are likely to have a BMI of over 16.

USRC discloses the following data:

> 40% of individuals that have been on Special Assistance Benefit for less than two years are likely to have a BMI of over 16.

Figure 3: non-identifiable statistical information

This is cannot be personal data in any organisation's hands, even USRC's, because the data has been irreversibly aggregated in such a way that even with USRC's 'key', this information does not *relate* to a particular, identifiable individual. This illustrates an important difference between aggregated and individual-level data.

Alternatively, USRC could have disclosed the following data in response to a different request:

> One individual of a cohort of five in the study had a BMI of over 16 having claimed Special Assistance Benefit for over five years.

Figure 4: potentially identifiable statistical information

Even though this information relates to only one individual, it is still not personal data once disclosed, provided that no other organisation knows the identity of the individuals taking part in USRC's study - or has the other information needed to link this

information to a particular individual. However, this information would still be personal data in the hands of USRC because its researchers could, if they wanted to, use the other information they have, to find out that this information relates to Mrs E Seniuk. No other organisation could do this unless they have the additional information needed to link the information to her.

We know that USRC could combine the information in *Figure 4* with other information it has in order to identify Mrs E Seniuk. However, the information above does not, in itself, constitute personal data because no one can be identified from just that information, except when it is in the hands of USRC. Once it comes into the possession of an organisation that does not hold the 'key' information, nor is likely to hold it (because USRC keeps the 'key' secure), it ceases to be personal data.

However, someone who knows Mrs E Seniuk – for example a family member – and knows that she took part in the research and that she has claimed Special Assistance benefit for over five years might now be able to discover that she has a BMI of over 16. However, this does not mean that the information in itself constitutes personal data about Mrs E Seniuk – except when it's in the hands of USRC. Releasing the information may, though, present a privacy risk, albeit a minor one because anyone capable of deducing Mrs E Seniuk' identity would already have to have a great deal of knowledge about her – none of which is in the public domain.

Barnardisation and 'blurring'

There are various methods of 'blurring', disguising or systematically altering data to reduce the risk, or make it less likely or impossible, for a link to be established between statistical information and other information that identifies a particular individual. It is particularly relevant to the sort of information in *Figure 4*. A 'blurred' version of *Figure 4* might look like this:

> 0-3 individuals of a cohort of 3-7 in the study had a BMI of between 15 and 17 having claimed Special Assistance Benefit for 4–6 years.

Figure 5 – 'blurred' statistical information

'Blurring' the information in this way means that no-one, not even Mrs E Seniuk herself could say "that information is about me". It is certainly possible to use techniques like this so that even

the original data controller can no longer establish a reliable link between the disclosure-controlled information and the individuals whose personal data this information was derived from.

Re-identification risk

USRC could use the following key to 're-identify' *Figure 2* type information.

Name, address: Mr B Stevens 46 Water St Stevenham	Research cohort ref. no. = 1A5

Figure 6: a re-identification key

Any organisation with the information in Figure 2 and access to this 'key' would clearly be able to discover that Mr B Stevens has been on Special Assistance benefit for less than 2 years and has a BMI of 15. However, the re-identification process is only possible here because the information in Figure 2 is divided into separate data fields that relate to a particular individual, allowing other information to be combined with it, resulting in 're-identification'. The process would not be possible where the de-personalised information is no longer separated into fields that relate to a particular individual. This might be the case where aggregated information is derived from the set of personal data.

Reference numbers and identification

There is a significant difference between USRC releasing this information:

1. Research cohort reference no.	2. Period of Special Assistance benefit	3. Body mass index	4. IB / BMI correlation score	5. Age range
1A5	<2 years	15	High	40-45

Figure 7

and this:

1. National insurance no.	2. Period of Special Assistance benefit	3. Body mass index	4. IB / BMI correlation score	5. Age range
NA111213Z	<2 years	15	High	40-45

Figure 8

The difference is that the data in Figure 7 only contains the '1A5' reference number – this is allocated by USRC for its own purposes and the 'key' linking it to Mr B Stevens is held securely by USRC and is never disclosed. Another organisation could use 'IA5' as an identifier – for example to match information about the same individual over time, for example, to monitor changes to Mr B Stevens' BMI in a longitudinal study. However, no organisation could use this reference number to link the information in Figure 7 to any other information about the same person, provided the 'key' is kept secure. Nor could '1A5' be used to take any action in respect of Mr B Stevens – for example to contact him to offer health advice.

However, many organisations – all employers for example - hold National Insurance numbers meaning that it is far more likely that the information in Figure 8 would be matched with other information to identify Mr B Stevens explicitly and to take action in respect of him. For example, an employer could, if he or she so wanted, take the NI number from Figure 8, check it against its own records, discover that the number relates to one of its own employees and offer Mr B Stevens occupational health advice. Although this is an unlikely scenario, it is certainly possible, and illustrates the difference between a '1A5' type number and one that is in wider circulation, such as an NHS or NI number. It also shows why a person's name and address is such a powerful identifier; because the same information is held and used by so many different organisations.

Some reference numbers are derived from, or may include, biographical information about individuals. For example, USRC could have allocated the following research cohort reference number to Mr B Stevens: 'BS2004169'. This is made up of his initials, date of birth and a check number. It is fairly likely that another individual or an organisation could deduce that the reference number relates to Mr B Stevens and find out from Figure 2 type data that his BMI is 15. For example, Mr B Stevens' employer or GP could probably do this if motivated to do so. This illustrates that an identifier that is formed from other biographical data that is relatively widely held – eg someone's date of birth – carries a relatively high risk of re-identification through matching it against other information sources.

Data-matching using unique patterns

It can be possible to determine that one piece of information relates to the same person as another, even though the information contains no unique identifiers, such as a reference number or name and address.

For example, the following data shows detailed fluctuations in an individual's BMI during the first six months of 2011 and then for the whole year:

An individual's BMI fluctuation:
14.1 13.9 13.4 13.2 13.1 13.1
14.1 13.9 13.4 13.2 13.1 13.1 13.2 13.5 13.4 13.4 13.8 14.0

Figure 9: a unique pattern

Even though this information contains no identifiers at all, it would certainly be possible for anyone to deduce, with a very high degree of certainty, that the second string of information relates to the same person as the first, even though the data is released in two batches. However, this does not mean that an individual has been identified or, therefore, that personal data has been disclosed.

What does it mean to identify someone?

NRC (the other research organisation) holds *Figure 9*-type detailed BMI information relating to all the members of the research cohort:

BMI fluctuation data: 2011

J	F	M	A	M	J	J	A	S	O	N	D
15.2	15.1	15.0	14.8	14.7	14.6	14.9	14.7	14.8	14.7	14.4	14.8
14.1	13.9	13.4	13.2	13.1	13.1	13.2	13.5	13.4	13.4	13.8	14.0
16.1	16.2	16.4	16.8	17.0	17.1	17.2	17.9	18.4	18.4	18.2	18.0
18.9	18.7	18.9	18.7	18.8	18.9	19.0	18.9	18.8	18.8	18.8	18.9
20.5	20.4	20.3	20.2	20.1	20.0	19.9	19.8	19.7	19.6	19.5	19.3

Figure 10 – a set of 'unique pattern' information.

NRC took this information from the Digest of Public Health's website, which routinely publishes data sets for use by medical researchers and others. The data was provided to the Digest by USRC.

The *Figure 10* data-set clearly *relates* to five individuals. This does not mean though that NRC can *identify* any living individual from that data. NRC has no other information in its possession that allows identification, nor is it likely that the other information needed to allow identification will come into NRC's – or any other organisation's possession - because only USRC holds that data and its research protocols specifically prohibit its disclosure.

However, USRC then releases some additional data and NRC downloads it from the Journal's website. It shows BMI fluctuation data from the last quarter of 2011 and the first quarter of 2012:

BMI fluctuation data: 10-2011 – 3-2012

O	N	D	J	F	M
14.7	14.4	14.8	14.8	14.6	14.8
13.4	13.8	14.0	14.2	14.1	14.0
18.4	18.2	18.0	18.3	18.5	18.5
18.8	18.8	18.9	18.8	18.8	18.8
19.6	19.5	19.3	19.2	19.1	19.0

Figure 11 – additional 'unique pattern' data

It would be possible for NRC to match the Figure 10 and Figure 11 data to deduce, with a high level of statistical certainty, that the first research subject, who had a BMI of 14.8 in April 2011 had a BMI of 14.8 in March 2012 – a piece of information not available from either data-set

However, even if NRC carries out a matching exercise using the Figure 10 and Figure 11 data, it only results in this:

BMI fluctuation data: 1-2011 3-2012

J	F	M	A	M	J	J	A	S	O	N	D	J	F	M
15.2	15.1	15.0	14.8	14.7	14.6	14.9	14.7	14.8	14.7	14.4	14.8	14.8	14.6	14.8
14.1	13.9	13.4	13.2	13.1	13.1	13.2	13.5	13.4	13.4	13.8	14.0	14.2	14.1	14.0
16.1	16.2	16.4	16.8	17.0	17.1	17.2	17.9	18.4	18.4	18.2	18.0	18.3	18.5	18.5
18.9	18.7	18.9	18.7	18.8	18.9	19.0	18.9	18.8	18.8	18.8	18.9	18.8	18.8	18.8
20.5	20.4	20.3	20.2	20.1	20.0	19.9	19.8	19.7	19.6	19.5	19.3	19.2	19.1	19.0

Figure 12 – 'matched' unique pattern data

Again, whilst this dataset contains detailed information that relates to various individuals over a period time, it still does not identify any individual, and cannot be used to do so unless the 'key' information needed to facilitate re-identification is disclosed – and USRC's procedures are designed to ensure that this will not happen.

Annex 2 – Anonymisation case-studies

Case study 1: limited access to pharmaceutical data

In a clinical study, only key-coded data is reported by clinical investigators (healthcare professionals) to the pharmaceutical companies sponsoring the research. No personal data is disclosed. The decryption keys are held at study sites by the clinical investigators, who are prohibited under obligations of good clinical practice and professional confidentiality from revealing research subject identities. The sponsors of the research may share the key-coded data with affiliates overseas, scientific collaborators, and health regulatory authorities around the world. In all cases, however, recipients of the data are bound by obligations of confidentiality and restrictions on re-use and re-identification, whether imposed by contract or required by law. Given these safeguards, the risk of re-identification of the key-coded data disclosed by a pharmaceutical sponsor to a third party under such obligations is extremely low.



Case study 2: using mobile phone data to study road traffic speeds

This shows how potentially quite intrusive information – in this case geo-location information derived from mobile phone data - can be converted into an anonymised form and used safely for a quite different purpose.

A telecommunications provider has a large list of subscriber records. Each contains:

- a mobile phone number
- an approximate location
- a date and time.

Some of the location data will relate to phones in cars travelling on the roads. The company wants to release a data set to a research body that will analyse it to derive information about traffic speeds on the roads - by calculating how fast individual phones are moving between particular locations.

To reduce the personal data content of the data set, the company replaces the phone numbers by dummy values. If the company just removes the phone numbers, clearly all the desired value is lost. If the company aggregates the numbers, the valuable information content will be significantly reduced. If the company randomly replaces every individual instance of a phone number, records that were linked in the original set will not be linked in the modified set, so again the required value will be lost. Instead, the company makes sure that the same real phone number is always replaced by the same dummy phone number, so that related records can still be linked.

The company can do this by:

- **Encryption** of the individual data records, eg by using the AES encryption algorithm - not a probabilistic encryption algorithm. As well as ensuring that identical original values are always mapped to identical modified values, this also ensures that different original values are always mapped to different modified values - there are no accidental "collisions". It is essential, of course, to keep the cryptographic key secret.
- **Tokenisation**. This means creating a mapping table, which maps values in the original data set to modified values. When producing the modified data set, as the company works through each input value in turn:

- if this input value already exists in the table, the output value indicated by the table is used;
- otherwise the company creates a new table entry for this input value, with the output value selected randomly subject to the constraint that the company never uses a value it has used before.

In this case it is essential that the mapping table is kept secret - it becomes the equivalent of the cryptographic key.

- **Randomisation** without guaranteeing uniqueness. In effect this is the same as tokenisation but without the constraint that a newly selected output value must be different from any that has been used before. (It may be that collisions do not matter much, or if the set of possible output values is very large then accidental collisions may be so improbable that will not be a problem.)

With either encryption or tokenisation, if the owner of the original data set retains the cryptographic key or the mapping table, then it may be able to translate analysis carried out on the modified data set back into results on the original data set. This can be very valuable for some applications.

Case study 3: analysing passengers' journey times

A public transport company uses its Go-Card data to carry out a study showing the amount of time commuters of particular ages take to make various journeys. It then uses anonymised data for accessibility planning purposes. The nature of data here is such that techniques such as pseudonymisation, hashing and data banding can be used to anonymise the data effectively:

Go-card no.	Passenger DoB	Start point	End point	Journey time
WT98765G	01/09/1973	Brooks End	Tree Street	17m 45s
WT45678B	18/09/1933	Brooks End	Tree Street	15m 05s

and this:

Hashed* passenger ref. no.	Age band	Start Point	End Point	Journey time
14793X…	35 - 45	Brooks End	Tree Street	18m
23955P…	75 – 80	Brooks End	Tree Street	15m

* a keyed cryptographic hash function such as SHA356

Case study 4: publicly available information and anonymisation risk

An electoral register entry includes the name and address of those eligible to vote. It also contains the dates of birth of those approaching voting age.

Ms K L Thomas: 1 Sandwich Avenue, Stevenham, SV3 9LK.

The public availability of the electoral register means that it would be easy to link Mrs Thomas to information about her property – for example its 'sold for' price on a property website.

Had Ms Thomas been a 17 year old, the publication of her date of birth might also present a re-identification threat, where, for example, an 'anonymised' research database is published that contains the years and months of birth and partial postcodes of research subjects.



Case study 5: a summary of a freedom of information decision notice relating to the disclosure of personal data. The full decision notice – and many others – are available on the ICO website.

Police: crime statistics at street level (adapted from decision notice FS50161581)

In 2007 a police force received a request for the number of burglaries in two specific streets, Daisy Lane and Iris Drive, over the years from 2004-2006. The police force refused the request under section 40 of the Freedom of Information Act 2000 (FOIA). In short, this exempts personal data from disclosure where disclosure would contravene the data protection principles.

Following a complaint to the ICO, the Information Commissioner considered whether the police force had acted in accordance with FOIA in withholding the requested information.

The Commissioner examined the statistics and took the view that it was not possible to identify any individual from the statistics alone. The Commissioner also considered the statistics in the context of other factors and information which might lead to the identification of an individual. There were 13 properties in Daisy Lane and 83 properties in Iris Drive. The Commissioner noted that the number of properties falling within the area that that the statistics related to was relatively small. Despite this, the Commissioner still took the view that the requested information would not lead to the identification of any individual.

The police force had argued that individuals with local knowledge would be able to identify individuals from the information. However, the Commissioner found no evidence to suggest that disclosure of the requested statistics would lead to any individual's identification. Therefore the Commissioner concluded that the statistics were not personal data and could not be withheld on s.40 grounds.

This case illustrates the need for careful judgement based on the circumstances of each case. Had the number of properties or the way the statistics were complied been different, the Commissioner may well have agreed with the public authority that section 40 was engaged. The case also illustrates that in some circumstances, for identification to take place the 'intruder' would already have a great deal of knowledge – in this case that a particular property or individual had been associated with a particular crime.

Case study 6: anonymising qualitative data

This shows how a piece of qualitative personal data – in this case an interview with a child - can be converted into an anonymised form which still contains valuable information but does not identify the child.

Original text

Interview recorded: 3pm, 10 October 2011
Interviewee: Julius Smith
DoB: 9 September 2005
School: Green Lanes Primary School

I live on Clementine Lane so I walk to school every day. I live in a flat with my parents and my Uncle Jermaine. When I get home from school I watch TV. I don't like reading but I like watching Harry Potter films. My favourite subject at school is art. My teacher is Mr Haines and he is very nice. I used to get bullied by Neil and Chris but I told Mr Haines and they stopped.

I play football for Junior Champs, and we are good. I play midfield.

Anonymised text

Interview recorded: October 2011
Interviewee ref: 2011/67
School year: Key Stage 1
School local authority area: Lynenham District Council

I live in [LM51 postcode] so I walk to school every day. I live with [family members]. When I get home from school I watch TV. I don't like reading but I like watching Harry Potter films. My favourite subject at school is art. My teacher is Mr [teacher's name] and he is very nice. I used to get bullied by [other pupils] but I told [the teacher] and they stopped.

I play football for [a local team], and we are good. I play midfield.

Case study 7: this shows the importance of third parties' prior knowledge in assessing re-identification risk and illustrates some means of reducing this risk

An engineering firm is carrying out a study of its employees' exposure to dermatitis-causing chemicals

Human Resources summary employee record:

Employee name:	F Gradwell
DoB:	01/09/1973
Sex:	M
Address:	16 Tree Street, Stevenham, SV8 6QP
Start date:	11/06/1992

Anonymised research database extract:

Age:	39
Sex:	M
Postcode:	SV8 6QP
Period of service:	20 years 5 months
Contact dermatitis:	Positive

It is unlikely that the firm's HR department would make its HR summary employee record publicly available. However, Mr Gradwell's workmates, friends and family members may well know his date of birth, address and (approximate) start date. Mr Gradwell might post this information on the internet himself as part of his social media profile.

This means that if someone – 'a motivated intruder' – wanted to, they could combine the datasets together to deduce with a fair degree of reliability that Mr Gradwell has dermatitis. This would be made much easier if it was known that the research data – published by a local university – relates to the employees of the firm where Mr Gradwell is known to work.

Employee name:	F Gradwell
DoB:	01/09/1973
Sex:	M
Address:	16 Tree Street, Stevenham, SV8 6QP
Start date:	11/06/1992
Contact dermatitis:	Positive

The risk of identifying Mr Gradwell as a dermatitis sufferer would be reduced if the data was 'blurred' in the following way:

Age range:	35-45
Sex:	M
Location:	Stevenham
Period of service:	18 – 22 years
Contact dermatitis:	Positive

It is unlikely that anyone – even someone who knows Mr Gradwell – could identify him and find out that he has dermatitis from this data set.

The risk would be reduced further if the data was presented in a non individual-level form:

Stevenham branch: 15% of male employees with 18 – 22 years' service have contracted dermatitis.

Case study 8: customers' purchasing habits - linking anonymised data

BuySome.com analyses its customers' purchasing habits to target relevant special offers at them. To do this, its systems analyse information in a personally identifiable form and send out vouchers to shoppers using its loyalty card database of names and addresses.

BuySome has been asked to take part in a research initiative run by a third party. This will involve correlating shoppers' purchasing habits with public health data about diabetes rates. Each organisation will use an extract from a sample of a group of individuals' health information and purchasing data.

In order to do this BuySome and other local supermarkets use a secure-keyed cryptographic hash to generate unique reference numbers from customers' names and addresses. GP surgeries use the same algorithm to generate unique reference numbers from their patients' details. Once the reference numbers have been created, both BuySome and the GP surgeries delete the key used in the hashing process.

This results in two anonymised datasets that the researchers can match together and analyse even though they cannot identify any individual. The researchers add another round of encryption to ensure that neither the participating GP surgeries nor BuySome could ever link the data back to individual patients' or shoppers' identities.

Case study 9: customer analytics

A North American home-ware retailer ("HW") was experiencing declining sales. In order to address this situation, HW needed to better understand its customers' requirements so it could improve its sales.

HW identified that the analysis of historical point of sale transactional data (POS data) would enable it to better understand what customers were buying in HW stores. HW engaged a third party, Research Direct, to help undertake an analysis of its POS data. Due to payment card regulations, HW was prohibited from sending raw POS data (which included credit card payment details) to Research Direct. In order to comply with the payment card regulations, HW applied one way encryption to the credit card data (contained in the POS data) for the purposes of transferring data to Research Direct. Research Direct was then able to analyse, over time, purchases made using the same payment card (using the encrypted key) and therefore enriched HW's understanding of its customers through analysis, like customer segmentations. Top-level findings were then shared with HW.

By using this method to anonymise its data, HW was able to analyse 82% of sales (the remaining 18% were cash purchases).

This enabled HW (a formerly struggling retailer) to accurately analyse its customers' motivations when buying products and specifically what they bought over time.

Case study 10: suppression Rules Applied to Data for the Longitudinal Study of Young People in England

This shows how anonymisation techniques can be used by a government department to protect the identities of respondents to a survey on bullying, educational attainment and other issues pertinent to young people.

Suppression at Wave 7 was informed largely by the Office of National Statistics (ONS) published guidance which sets out the GSS Microdata Policy for Social Surveys (http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-social-survey-microdata/index.html). Very few rules from this guidance are specific so a certain amount of interpretation had to be applied.

Different from previous waves, it was felt that variables should not be recoded for the sole purpose of avoiding suppression. This decision was made partly to reduce the likelihood of misinterpretation of recoded variables away from the original questionnaire, but also because there was less resource available for enhancement at this wave, as the work wasn't contracted out this year.

Of the 932 variables which were in the original dataset, 425 remained after suppression had been completed. All suppressed variables will still remain available to data customers via the Longitudinal Surveys Team, but only after completing the LSYPE Confidentiality Agreement. The proportion of variables suppressed is believed to be higher than at previous waves, However, this is not a result of over-sensitivity. As the respondents got older and the sample size reduced through attrition, questions relating to less common activities at age 19 (such as apprenticeships and non-HE qualifications) have seen fewer overall responses and are therefore more vulnerable to suppression. The policy to not recode or band variables solely for suppression reasons will also have impacted the number suppressed.

The rules by which variables were suppressed were as follows:

- **Sensitivity** – Some variables that are highly sensitive are suppressed purely because they are defined as such in the ONS guidance, whilst others are clearly sensitive regardless of whether they are mentioned in the guidance.
- ***Example variables***: 'sexuality', un-banded pay information, 'number of sexual partners', 'what bullying was motivated by', 'sexual orientation', 'marital status'.

- **Low Numbers** – Once the number of overall responses to a particular question drops to a low level, the question may become identifiable. For that reason all questions with less than 200 overall responses have been suppressed. In addition, where there are multiple responses to a question, it is possible that a response given by only a minority of respondents is also identifiable. For this reason, where less than 10 responses have been given to an answer that identifies something *factual*, all variables relating to that question have been suppressed. It should be noted that *attitudinal* questions are not bound by this rule, in addition to responses of 'Don't know', 'Refused', 'Other' or similar.
- ***Example variables***: 'Number of GCSEs studying for since September 2009', 'number of OCR qualifications studying for', 'reasons for doing apprenticeship', 'whether usually pays for childcare', 'types used - a nursery school or nursery class', 'how health problem/disability affects life', 'continence', 'type of bullying experienced in last 12 months during work/study/training', 'physical abuse'.
- **Identifiable Detail** – In some cases, responses to questions provide large amounts of detailed information that is not necessarily sensitive, but is at such a low level that it becomes identifiable. In many cases this information is duplicated through derived variables that were created to make analysis on these topics sensible, so source variables with a high number of different responses are suppressed.
- ***Example variables***: 'Number of current jobs', 'amount of hours usually worked each week', detailed information about HE subject and institution.
- **Others** – Some variables asking for detailed information on benefits have been historically suppressed and therefore continue to be at Wave 7. Additionally one less detailed benefits question was incorrectly asked during fieldwork and so has been replaced by a derived variable that better reflects the answer to the question. There was also one question – about methods used to pay for fees and living expenses other than grants and bursaries - where there were a low number of responses to answers that were essentially *incorrect* as they were supposed to be specifically excluded by respondents when answering the question. The two incorrect responses with low response rates have been suppressed with the rest of the question left unsuppressed.

Case study 11: cryptographic hash technique.

1. Let *D* be the personal data you wish to anonymise.
2. Let *K* be a secret key, known only to the data controller.
3. Choose a secure cryptographic hash function *H*. You must take care to ensure that the hash function you are using is secure as older algorithms, which may have been used previously, may no longer be secure due to exposed vulnerabilities and should therefore not be used.
4. Compute the hash of the sequence *K D*, i.e. *H*(*K D*). This will generate a unique value which can be used to replace the personal data in the anonymised release, while still providing (a) a unique identifier for the personal data, and (b) a means for the data controller to retrieve the original data, since he knows both *D* and *K*, he can easily re-compute or store *H*(*K D*) for each data point.

Due to the nature of cryptographic hashes, it is implausible to reverse the hash, even with knowledge of the value of *K*. However, if the domain of *D* is small, knowing both the hash value *H*(*K D*) and *K*, one might be able to guess *D* by computing *H(K D)* for each possible value of *D* and comparing this with the anonymised data set. For this reason it is important that *K* remains a secret known only to the data controller.

Note: If there is a likelihood that there are multiple records for the personal data *D* then the value of *H(K D)* will be the same for each of these records in the anonymised release. If you do not wish to show these relationships, add a "salt" value *S*, which is a randomly chosen value that you append both to the sequence that is hashed as well as the final identifier. So, rather than *H*(*K D*), compute *H*(*K S D*), and then use *SH*(*K S D*) as your full identifier. *S* should, ideally, be selected using a cryptographic strength random number generator. As before, *K* must remain a secret, known only to the data controller.

Annex 3 – Practical examples of some anonymisation techniques

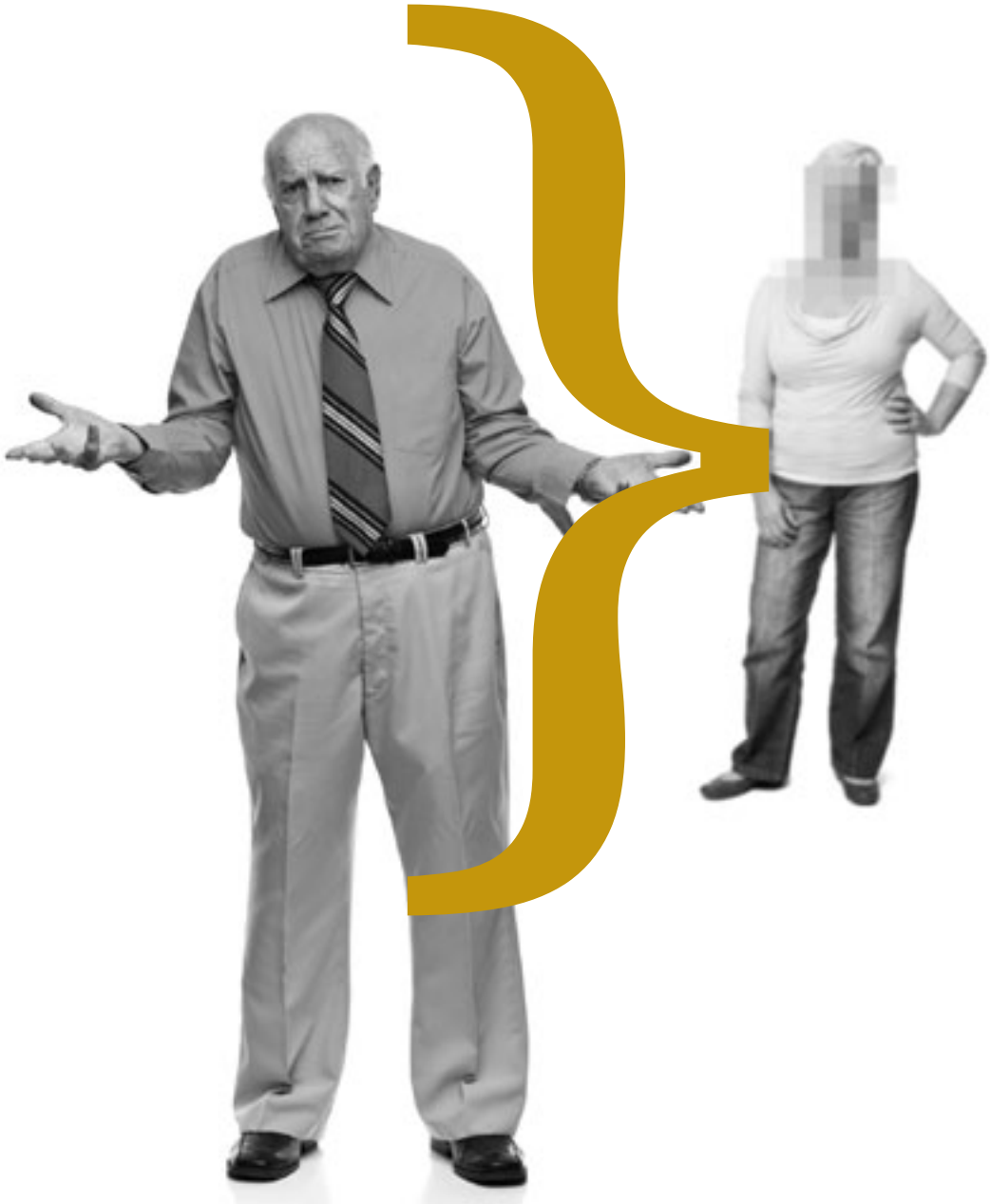
– drawn up by Mu Yang, Vladimiro Sassone and Kieron O'Hara
at the University of Southampton

In this annex, we will set out a few examples of the anonymisation of data, to indicate the range of techniques available to the information manager. The aim is not to provide a manual of anonymisation, but to give a flavour of the field, and of the variety of the options. We do not pretend that this is an exhaustive list of methods, or that the methods we have chosen are applicable to all anonymisation problems. We try to keep the language as accessible as possible; however, some of these techniques are statistically quite complex, which in some cases is inevitably reflected in the descriptions.

We focus on de-identification (whether complete or partial) in this annex, and look at the trade-off between preserving data utility and preserving anonymity. We do not discuss refinements of that task (such as pseudonymisation or deterministic modification). We concentrate on examples using quantitative or categorical data, as being somewhat more common and well-understood. There are several types of qualitative data which it may be desirable to anonymise, eg free text comments, geolocation data, or purchasing records – which we do not cover here, partly for reasons of space, partly because of the relative complexity (for more on anonymising qualitative data, see http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation?index=2).

We do not recommend particular techniques over and above others; the choice of a particular method of anonymisation will depend on many factors, including an understanding of the potential risk of exposing personal data inappropriately, the sensitivities of the data, and the amount of control that the data controller has over the uses to which the anonymised data will be put. It may also be that the application of certain techniques could severely damage the utility of the dataset in a particular context.

Hence the choice of an anonymisation technique should always be a matter for the data controller's judgement, based on the context of data sharing or use.



A – Data reduction

A.1 Removing variables

Description: A variable is a characteristic or attribute of an individual – for each individual the variable will have a value (eg the values of the variable NAME for the three authors of this appendix are Mu, Vladimiro and Kieron). The simplest method of anonymisation is the removal of variables which provide direct or indirect identifiers from the data file. These need not necessarily be names; a variable should be removed when it is highly identifying in the context of the data and no other protection methods can be applied.

A variable can also be removed if it is deemed too sensitive for public use or irrelevant for analytical purpose (eg if a dataset intended for reuse for market research purposes included a variable which expressed whether the individual has been convicted for a certain class of sexual offences, that variable could simply be removed as too sensitive). Of course, deciding what is 'sensitive' is an art rather than a science and will depend on context; such judgments are part of a data controller's risk assessment.

Example: If the intruder was personally acquainted with the group in example one, then the 'ethnic' variable could be identifying for a large fraction of the group members. If this variable was simply removed from the record, the identification risk falls dramatically.

Removing variables

Example one: the removal of direct identifiers

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month	Ethnic
22	F	SO17	£20,000	£1,100	British
25	M	SO18	£22,000	£1,300	Irish
30	M	SO16	£32,000	£1,800	African
35	F	SO17	£31,500	£2,000	Chinese
40	F	SO15	£68,000	£3,500	Pakistani
50	M	SO14	£28,000	£1,200	British

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month	Ethnic
22	F	SO17	£20,000	£1,100	British
25	M	SO18	£22,000	£1,300	Irish
30	M	SO16	£32,000	£1,800	African
35	F	SO17	£31,500	£2,000	Chinese
40	F	SO15	£68,000	£3,500	Pakistani
50	M	SO14	£28,000	£1,200	British

Comments: This technique is subject to much information loss if the variable is very important to the analysis.

A.2 Removing records

Description: Removing records of particular units or individuals can be adopted as an extreme measure of data protection when the unit is identifiable in spite of the application of other protection techniques.

Example: In example two, only one male is involved, so the intruder can easily identify him in the data if he/she is acquainted with the participants. Removing this record prevents his personal data from being recovered from the table.

Removing records

Example two: the removal of a particular record which is easy to identify

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
~~30~~	~~M~~	~~SO16~~	~~£32,000~~	~~£1,800~~
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Comments: Removing records is similarly damaging to the information content of the matrix to removing variables; the latter removes a column from the table, while the former removes a row. In this example, it has been deemed preferable to remove this particular record rather than removing the variable 'Gender' from all records. However, removing records will significantly impact (ie distort) the statistical properties of the released data. And the risks of jigsaw identification may grow where aggregated data are also published as the removed records could be inferred by subtracting the published records from the aggregation. In example two, if the total income of the group was published, that would enable the disclosive information (ie information that, even if not identifying, reveals hitherto unknown information about a known individual) to be inferred (as the sum of the incomes of the published records is £169,500, it can be subtracted from the aggregated income total of £210,500 to produce the one male's salary).



A.3 Global recoding

Description: This method makes variable values less specific, and the table correspondingly less informative. For a categorical variable (ie one that categorises the units), several categories are combined to form new (less specific) categories, thus resulting in a new variable. A continuous variable is replaced by another variable which aggregates ranges of the continuous variable. In other words, the global recoding method consists in aggregating the values observed in a variable into pre-defined classes. Every record in the table is recoded.

A more informative type of recoding involves recoding only the outliers (i.e. unusually high or unusually low values). For instance, incomes between, say £20,000 and £60,000 would be reproduced in the recoded table, but outside that range would be recoded as <£20,000 or >£60,000. This type of recoding leaves the vast majority of 'normal' values unchanged.

Example: In example three the variables 'Age' and 'Postcode' are aggregated into new classes, each of which has values as a range. The more specific values have unique mappings to a less specific range. We also recode the 'Income' and 'Expenses' variables into the classes low, medium and high, again using a unique mapping.

Global recoding

Example three: aggregating the values observed in variables into pre-defined classes

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17-19	low	low
25-29	M	SO17-19	low	low
30-34	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	M	SO14-16	medium	low

Comments: Global recoding involves information loss via loss of specificity. A related drawback is that a recoding that suitable for one set of records might be completely unsuitable for another set. For example, the categories of 'Age' variable may protect identities in one example, but may still be used to disclose information in another. There are also obvious limits; we cannot simply recode 'Female' and 'Male' as 'Female or Male' (this is tantamount to removing the variable entirely).

Recoding the outliers has two advantages. First of all, the unusual information may be identifying or disclosive by virtue of its unusualness, and that is made less specific. Conversely, the majority of the cases can safely remain untouched, because a 'normal' value will be shared by many and so is much less likely to be disclosive. Secondly, there will typically be few outliers, and so most of the original information in the dataset will be preserved intact.

A.4 Local suppression

Description: Local suppression consists of replacing the observed value of one or more variables in a certain record with a 'missing' value. This is particularly suitable with categorical key variables (a key variable is a variable that a researcher is particularly interested in). When combinations of such variables are problematic, local suppression consists of replacing an observed value with 'missing' or some other value which shows that the original value has been suppressed. The aim of the method is to reduce the information content of rare combinations. The result is an increase in the frequency count of records containing the modified combination.

Example: In example four, as the combination "Age=20-24, Gender=F" is unique, an intruder may identify this individual if the intruder has information about a young lady in the cohort. If the number of females in the dataset is high, we can suppress the variable 'Age' of this record and recode it as 'missing'.

Local suppression

Example four: replacing the observed value of one or more variables in a certain record with a missing value

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17	low	low
25-29	M	SO18	low	low
30-34	M	SO16	medium	medium
35-39	F	SO17	medium	medium
40-44	F	SO15	high	high
50-54	M	SO14	medium	low

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
missing	F	SO17	low	low
25-29	M	SO18	low	low
30-34	M	SO16	medium	medium
35-39	F	SO17	medium	medium
40-44	F	SO15	high	high
50-54	M	SO14	medium	low

Unique combination

Comments: Local suppression should be applied only to records that contain combinations at risk. Once the local suppression technique is used, analysis of the data is not simple in the absence of highly specific software. Once more it can be easier to deduce the missing values if aggregated totals are also given. Furthermore, local suppression only works when there is sufficient variety to prevent the missing value being inferred (most obviously, if the categories used for 'sex' were 'male' and 'missing', it would not be hard to infer the sex of everyone).

Perturbing data involves changing some data values according to a set of principles. The aim is to disguise the records of individuals while leaving some wider properties of the population (eg mean, or average, values of the variables) unchanged. For instance, one relatively straightforward method of perturbation is Barnardisation, which involves adding or subtracting a constant from some values of some variables. In a Barnardised dataset, it is impossible to be sure (without supplementary information) which data is accurate, but the population statistics remain reliable. In this section, we discuss a number of more complex methods of perturbation.

Note that all the methods in this section may render the data unusable for research that relies on individual-level data as the data in the original datasets is perturbed.

B – Data perturbation

B.1 Micro-aggregation

Description: The idea of micro-aggregation is to replace an observed value with the average computed on a small group of units. The units belonging to the same group will be represented in the released file by the same value. The groups contain a minimum predefined number *k* of units. Here *k* is a threshold value and the partition is called a *k-partition*. In order to obtain micro-aggregates from a dataset with a certain number of records, these records are combined (usually in a meaningful order, such as size order) to form groups of size at least *k*. We do this by computing the average value of the target variable over each group and then replacing the original values with this average value. The mean value for the whole population remains unchanged.

So, for example, if we had 100 individuals in the dataset and wished to form a 4-partition, then segment the dataset into 25 groups of 4. For each group, the average value of the variable is computed, and that average replaces the observed value in the dataset. If a group of 4 individuals had ages 31, 33, 33 and 34, the age for each individual in the published dataset would be 32.75.

Micro-aggregation can be independently applied to one variable or a set of variables. In the former case, for different variables the dataset could be partitioned in different ways, so that an individual might not find itself in the same partition for different variables. It is then called *individual ranking*. In the latter case, then a number of groups would be formed, and the average of several variables computed in each group. When all the variables are averaged at the same time for each group, the method is called *multivariate micro-aggregation*.

Example: In example five, the intruder may identify some individual if he has information about their incomes. So if this is a real danger, we apply micro-aggregation to the variable 'Income'. We firstly sort the values from small to big, and then perform a (i.e. we set *k* to). So the group number *g* in this small example of 6 individuals is 6 ÷ 3 = 2. Then we compute the average value for each group and replace the original value by the average value.

Example five: replacing an observed value with the average computed on a small group of units then the units belonging to the same group will be represented by the same value,

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
50	M	S014	£28,000	£1,200
35	F	SO17	£31,500	£2,000
30	M	SO16	£32,000	£1,800
40	F	S0O15	£68,000	£3,500

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£23,333	£1,100
25	M	SO18	£23,333	£1,300
50	M	S014	£23,333	£1,200
35	F	SO17	£43,833	£2,000
30	M	SO16	£43,833	£1,800
40	F	S0O15	£43,833	£3,500

k partition = 3

Comments: This method guarantees that at least units in the file are identical; the information loss about specific individuals is high.

B.2 Data swapping

Description: Data swapping alters records in the data by switching values of variables across pairs of records in a fraction of the original data. The purpose is to introduce uncertainty for a data user or intruder as to whether records correspond to real data elements.

The variables that will be swapped are called *swapped attributes* or *swapping attributes* and the fraction of the total *n* records in the microdata that are initially marked to be swapped is called the *swap rate*, and is denoted by *r*. Typically, *r* is of the order of 1-10% (so that the fraction of attributes swapped will usually be less than one in ten).

In some situations there may be conditions on what pairs of records can be swapped. These conditions place constraints on the variables in order for one record in the pair to be a *feasible* swap candidate for the other. Such variables whose values define the feasibility of swap candidates are called *constraining attributes*. For example, one might only want to swap (say) salary values for two individuals if they are located in the *same* postcode. This is to ensure that the average salary for each postcode remains unchanged by the data swapping; the postcode is the constraining attribute. In that case if two individuals live in different postcodes, then their salary values cannot be swapped. As another type of example, one could swap salary values for two individuals only if they are of *different* sexes; the reason behind this might be to reduce the amount of information that could be deduced from personal knowledge of the individuals involved. Therefore, when swapping is applied, the necessary parameters are: the swapped attributes, constraining attributes and swapping rate.

Example: In example six, the first and fourth records are more vulnerable to attack as their variable 'Age' has unique values: '20-24' and '35-39' respectively, unlike the rest of the population. We designate 'Age' as the swapping attribute, and also set 'Gender' as a constraining attribute, thereby allowing swaps of Age only between those records with the same value of variable 'Gender'. In this example, the swapping rate r = 2 ÷ 6 = 33.3%. The high value of the swapping rate is of course due to the small population in the example.

Data swapping

Example six: altering a proportion of the records by swapping values of a subset of variables between selected pairs of records (swap pairs).

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17-19	low	low
25-29	M	SO17-19	low	low
30-34	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	M	SO14-16	medium	low

value unique

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
35-39	F	SO17-19	low	low
25-29	M	SO17-19	low	low
25-29	M	SO14-16	medium	medium
20-24	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	F	SO14-16	medium	low

Swapped attribute is Age.

Swapping rate: r=33.3%*.

Constraints: only allow swaps of Age between records with the same value of Gender

* The rate r is typically in the range of 1-10%. We choose 33.3% because of the limited number of records

Comments: Swapping does not change the distribution of any variable, but still there is the anonymisation trade-off that lowering the risk implies higher information loss.

B.3 Post-Randomisation Method (PRAM)

Description: The Post-Randomisation Method is a probabilistic method to perturb categorical variables. In the released file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a probability mechanism called a *Markov matrix*. This is quite a complex method, which is somewhat difficult to describe in straightforward language.

Suppose we have a categorical variable *V* which we wish to perturb, and suppose that that variable has *K* categories (so, for example, 'sex' is a categorical variable with two categories). For that variable *V*, we can decide to change one of the *K* values to another with a certain probability fixed in advance; we can arrange these probabilities in a *K* x *K* matrix (the Markov matrix), where, say, the second cell in the fourth row is the transition probability that when we have a value in the fourth category in the observed data, we transform it into the value of the second category in the published data. We can then decide to transform or perturb the data or not, depending on a random process. So, for instance, if our categorical variable was 'sex', and all the probabilities in the 2 x 2 Markov matrix were 0.5, we could toss a coin each time to decide whether or not to alter the attribution of M or F to each individual in the data.

In more detail, we begin with our categorical variable *V*. Let's call the same variable in the perturbed file *X*. Suppose also that these variables have *K* categories, which we can number from 1 to *K*. We define *transition probabilities* for each pair of categories from *V* and *X*; we denote the probability that, for *k* and *l* between 1 and *K*, when the value of the original variable *V* is *k*, it is transformed into the value *l* in the *X* variable in the perturbed file. The complete set of transition probabilities between all pairs of categories of *V* and *X* gives us a *K* x *K* matrix which is the Markov Matrix. The individual entries in the Markov Matrix are referred to as *p11*, *p12*, *p13*, *p21*, *p31*, etc, so that, say, *p31* is the probability that category 3 of variable *V* will be transformed into category 1 of variable *X* in the perturbed file. The general case, the probability of transforming *k* into *l* is referred to as *pkl*.

Applying the matrix to the data then means that for each value *k* of *V*, the probability of the corresponding value of *X* in the perturbed data file is drawn from the probability distribution *pk1 ... pkK*. For each record in the original file, this procedure is performed independently of all other records.

Example: In example seven, suppose that the variable *V* is Gender with scores *V = 1* if male and *V = 2* if female. Applying PRAM with *p11 = p22 = 0.9* on the original dataset with three males and three females, would yield a perturbed file with the expected totals of three males and three females. In these records, one of these three 'males' was actually male and similarly, one of these 'females' was actually male.

Post-Randomisation Method (PRAM)

Example seven: producing a microdata file in which the scores on some categorical variables for certain records in the original file are changed into a different score according to a prescribed probability mechanism

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	M	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	F	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

target variable = Gender, the PRAM-matrix: p11=p22=0.9, p12=p21=0.1

Comments: Since PRAM uses a probability mechanism, an intruder can never be sure that a record describes the identified person whom the intruder thinks he has identified. There is a certain probability this has been a perturbed record. However, if the Markov matrix that is used when applying PRAM is known, the true data may be estimated from the perturbed data file.

B.4 Adding noise

Description: Adding noise, a method applied to numerical data, consists of adding a random value ε to all values in the variable to be protected. The distribution of ε has mean zero and predefined variance σ^2. In other words, the expected value of ε is zero (sometimes the value will be positive, sometimes negative), so that given that noise is added to enough values the additions will cancel themselves out, leaving the mean of the distribution unchanged. The variance defines the range of the additional ε; a small variance means that ε is unlikely to be very far from 0 (and so the numerical change in the data unlikely to be large in any instance), while a larger variance will allow greater perturbations of individual data values. This type of distribution, a *normal distribution*, is the most standard type of distribution in statistics, very well-understood and often encountered in practice with real-world data.

Example: In example eight, we apply this method on the variable 'Income' by adding noise values generated by a standard normal distribution.

Comments: This method is less effective if there are large differences between values, or there are some outliers. For example, in this example, if an intruder knows that exactly one individual has a much higher income than the others, he or she can still identify this individual in the perturbed file, and even make a reasonable guess at a plausible range for the individual's income.

Adding noise

Example eight: adding a random value ε, with zero mean and predefined variance σ^2, to all values in the variable to be protected

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Standard Normal Distribution: mean=0, variance=1

-0.171932015
1.862281351
0.959896624
-2.543129085
-1.049088496
-0.308324388

x 1000

-£172
£1,862
£960
-£2,543
-£1,049
-£308

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£19,828	£1,100
25	M	SO18	£23,862	£1,300
30	M	SO16	£32,960	£1,800
35	F	SO17	£28,957	£2,000
40	F	SO15	£66,951	£3,500
50	M	SO14	£27,692	£1,200

B.5 Resampling

Description: Resampling is also designed for numerical data, and again requires understanding of statistical methods. It has three steps. First, we have to identify the way that the sensitive or key data variables vary across the whole population. This means deciding what the population will look like if put on a graph; typically, the answer will be a type of reasonably well-known type of distribution.

There are a number of what are called *probability distributions* or *probability density functions* (so-called because they allow us to estimate the probability of a variable having a particular random value); in B.4 we met the most common kind of density function, the normal distribution (where the variable can take any numerical value and will group around a central average), but there are others – for instance a Poisson distribution (where the variable is a positive whole number, 0, 1, 2 etc, and tends to group around an relatively small average and then tail off gradually as we go to infinity) or a Beta distribution (where the variable is a real number between a pair of limits). Each such distribution will be completely characterised by a small number of parameters (for example, the normal distribution is described by the mean and variance as hinted above, while the Poisson distribution is described by a single number which equals both the mean and the variance, and the Beta distribution is described by two so-called shape parameters which alter the shape of the curve). There are many types of distribution in statistics, most very specialised and complex. The details are not important for the purposes of this appendix; the reader basically needs to be aware that a population's properties can be estimated and described using these statistical terms.

So the first task is to estimate how a particular variable for the whole population is distributed, and to estimate the values of the relevant parameters for the population. Note that this estimation is for the whole population, not for just the population in the database (so, for example, we guess the average salary, and the way the salaries vary, for the population as a whole, not just for the people we have on the database).

The second step is to generate a distorted sample artificially which has the same parameter values as our estimate. The sample should be the same size as the database.

The third step is to replace the confidential data in the database with the distorted sample. So, in the salary example, if we have 100 lines in our dataset, and having decided how salaries vary across the population, we generate an artificial distribution of 100 salaries that has the same mean and variance as the estimate for the whole population. We then substitute those 100 artificially generated salaries for the 100 observed salaries in the database.

In many cases, to preserve correlations with other variables than the confidential one(s), the sample should also be ordered before mapping, so that the values of the sample are in the same order as the values of the database they replaced.

The resampling procedure creates datasets – the resample – which have the same, or nearly the same, empirical properties functions as the original survey data and thus permit statisticians to perform meaningful analyses.

Example: In example nine, we resample the two variables 'Income (Jan)' and 'Income (Feb)' together by using the RSXL add-ins tools for Excel. We can see the original and perturbed datasets have the same mean of the two-month incomes.

In the second version of the example, the generated samples are ordered before mapping and replacement on the original data, so that relationships between variables (eg the correlation between age and income) are preserved to some extent.



Resampling

Example nine: drawing with replacement t samples of n values from the original data, sorting the sample and averaging the sampled values

Income Individual-level dataset

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£20,000	£23,000
25	M	SO18	£22,000	£22,000
30	M	SO16	£32,000	£30,000
35	F	SO17	£31,500	£35,000
40	F	SO15	£68,000	£58,000
50	M	SO14	£28,000	£29,000

Income Individual-level dataset

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£58,000	£58,000
25	M	SO18	£22,000	£20,000
30	M	SO16	£29,000	£20,000
35	F	SO17	£20,000	£68,000
40	F	SO15	£22,000	£30,000
50	M	SO14	£20,000	£31,500

Income Individual-level dataset with ordered mapping

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£20,000	£20,000
25	M	SO18	£20,000	£20,000
30	M	SO16	£29,000	£31,500
35	F	SO17	£22,000	£58,000
40	F	SO15	£58,000	£68,000
50	M	SO14	£22,000	£30,000

Normal distributed resampling. Hypothesis testing: Null Hypothesis	Grand mean: £33, 208

Comments: Given that the original data are sampled from a very large population, estimating the probability density function of the variables is hard to achieve and verify, as sufficient information about the true distribution of data may not be available. The data will only sometimes follow a specific theoretical distribution such as those discussed above, which may make creating the distorted sample more difficult. Information about individuals is lost, and the correlations between variables will be affected.

C – Non-perturbation methods

C.1 Sampling

Description: Sampling is one of the non-perturbative methods in anonymisation techniques, suitable when the original data is in sufficient quantity to make a sample meaningful. Instead of publishing the original microdata file, we take a sample from it and publish that without identifiers. The resulting sample may contain information which is sensitive and which in other circumstances could be quite disclosive. However, because there is no way of knowing whether a particular individual's data is included in the sample, it is unlikely, though not impossible, that it would actually be disclosive.

Two common types of sampling are simple random sampling, where all possible subsets of specified size sample have an equal probability of selection, and Bernoulli sampling, where each record in the sample is selected independently with a certain probability.

The probability that a random sample preserves the basic statistical properties of the original dataset can be calculated.

Comments: Sampling is suitable for categorical microdata, but its adequacy for continuous microdata is less clear in a general disclosure scenario.

Unlikely to be disclosive, but for unusual or unique individuals it remains a possibility if someone is aware of their unique qualities. For example, if it is known that there is only one teenage amputee in a small town, then that combination of information can be looked for in the sample. If it has been sampled, then the dataset could be disclosive. However, if there were more – say five such people – the appearance of one in the dataset would require the intruder to gather further information before he could be confident he had tracked down his target.

C.2 Cross-tabulation of data

Description: When we have a table of data with two or more variables, we can create another table by tabulating the two variables against each other direction, in effect aggregating the data. The resulting table is called a contingency table. It can protect the confidentiality in microdata, especially for large numbers, and is non-perturbative.

Example: In example ten, we generate a contingency table by tabulating the two variables 'Gender' and 'Education Level'. The contingency table does not contain the individual information, that is, we are not sure of the first individual's educational attainment.

Cross-tabulation of data

Example ten: generating the contingency table which does not contain the individual information.

Record NO.	Gender	Education Level
1	F	Undergrad Degree
2	M	Grad Degree
3	M	Doctorate
4	F	Doctorate
5	F	Doctorate
6	M	Undergrad Degree

	Undergrad Degree	Grad Degree	Doctorate	Total
Female	1	0	2	3
Male	1	1	1	3
Total	2	1	3	6

Comments: It is a non-perturbative method. The confidentiality can be compromised if rare situations are revealed via very small numbers of cases, which in turn can be linked to an individual. For instance, from the contingency table, the number of females having a Grad degree is zero. So the attackers are sure that no female has a Grad degree.

Reference

In this survey, the authors are indebted to many comments from those who responded to the ICO's consultation on the Code of Practice, and to the account in Molla Hunegnaw, African Centre for Statistics, Confidentiality and Anonymization of Microdata.



Notes

If you would like to contact us please call 0303 123 1113

www.ico.org.uk

Information Commissioner's Office,
Wycliffe House, Water Lane,
Wilmslow, Cheshire SK9 5AF

November 2012

ico.
Information Commissioner's Office

Upholding information rights