

Draft Anonymisation code of practice

For consultation

31 May 2012 – 23 August 2012

Contents

1. About this code	3
2. Anonymisation and personal data	7
3. Ensuring the effectiveness of anonymisation	11
4. Do you need consent to produce or disclose anonymised information?	24
5. Personal data and spatial information	26
6. Withholding anonymised data	29
7. Publication and limited disclosure	31
8. Governance	34
9. The Data Protection Act research exemption	38
Appendix 1: Practical examples and commentary on some key anonymisation techniques	40
Appendix 2: The data protection principles	52
Appendix 3: Anonymisation techniques	53
Appendix 4: Glossary	55
Appendix 5: Further reading and sources of advice	58

1 About this code

This code explains the implications of anonymising personal data, and of disclosing data which has been anonymised, in terms of the requirements of the Data Protection Act 1998 (DPA). It provides good practice advice that will be relevant to all organisations that need to convert personal data into a form in which the individuals to whom it relates are no longer identifiable – anonymised data. It also contains a number of examples that illustrate some of the techniques that can be used to anonymise personal data. Many important issues concerning anonymisation have arisen in the context of the Freedom of Information Act 2000 (FOIA) and the Freedom of Information (Scotland) Act 2002, and many of the good practice recommendations and techniques contained in this code will be of relevance to freedom of information practitioners. The code also contains an explanation of the DPA's research exemption and its relevance to anonymisation.

The code is intended to demonstrate that the effective anonymisation of personal data is possible, desirable and can help society to ensure the availability of rich data resources whilst protecting individuals' privacy. Anonymisation is of particular relevance now, 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 be used as a barrier to prevent the anonymisation of personal data, given that anonymisation is ultimately intended to safeguard individuals' privacy.

We use the broad term 'anonymisation' to cover the various techniques that can be used to convert personal data into an anonymised form. This could be done for statistical reasons, but the approaches described in the code could also apply to other forms of output, such as anonymised case studies, maps or images. Various key anonymisation techniques and an assessment of their strengths and weaknesses are illustrated in Appendix 1.

Finally, we include a glossary of the various technical terms that appear in this document.

Who should use this code of practice?

Any data controller who is involved in the production or publication of anonymised information should use this code to help them to understand how to adopt good practice. Much of the good practice advice will be applicable to public, private and third sector organisations. Some parts of the code will be more relevant to some organisations than others. However, the majority of the code will apply to all instances of anonymisation regardless of its scale and context.

How the code can help

Adopting the good practice recommendations in this code will help you to anonymise personal data so that individuals' privacy is not compromised. It will also give you the confidence to publish anonymised information, furthering organisational transparency and allowing you to explain your organisation's performance to the public, for example.

The code will help you to identify the issues you need to consider when deciding how to anonymise personal data. It will help you to adopt the most appropriate means of anonymising the personal data you hold. It will also help you to assess any risk associated with producing – and particularly publishing – anonymised information.

In the event of the Information Commissioner investigating an issue arising from 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 been done with due technical and organisational rigour.

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;

For consultation

- 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 '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 of personal data; and
- reducing questions, complaints and disputes about your publication of information derived from personal data.

The 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 economic benefits that the availability of rich data resources can bring;
- public confidence that information 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 Data Protection Act 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.

For consultation

This code is the Information Commissioner's interpretation of the standards the DPA requires and how the anonymisation of personal data can further compliance. It 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

Why anonymise?

The main rationale for undertaking anonymisation is to protect individuals' privacy, whilst making available the data resources that activities such as research and planning rely on. It is clearly legitimate to use personal data for particular purposes where the intention is to inform decisions about particular individuals, or to provide services to them, for example. However, where this is not the intention, then the objective should be to use anonymised information.

The Data Protection Act (DPA) is concerned with 'personal data'. 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. It follows therefore that information or a combination of information, that does not relate to and identify an individual, is not personal data and that the DPA does not apply to it. Clearly, effective anonymisation depends on a sound understanding of what constitutes personal data. See the ICO's technical guidance on '[Determining what is personal data](#)'¹.

The most explicit reference to anonymisation in European data protection law is in Recital 26 of the Data Protection Directive (95/46/EC). This 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. Recital 26 also 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. Recital 26 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 – therefore the examples and commentary in Appendix 1 should be particularly useful.

1

http://www.ico.gov.uk/for_organisations/guidance_index/~/_media/documents/library/Data_Protection/Detailed_specialist_guides/PERSONAL_DATA_FLOWCHART_V1_WITH_PREFACE001.ashx

The benefits of anonymisation

All organisations that process personal data are required by law to protect it from inappropriate disclosure. By 'disclosure' we mean providing personal data, or making it available, to a third party. However, these organisations may want or be required to publish information derived from the personal data they hold. For example, a health service organisation will be 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 (see Appendix 2). The principles are intended to protect the privacy of 'data subjects' – the individuals whose personal data relate to and identify. Importantly, the principles regulate the disclosure of personal data, and in some circumstances can prevent this. This means that, in general, it is preferable to disclose or publish anonymised information than personal data as fewer legal restrictions will apply. It is also far easier to use anonymised information in new and different ways because the DPA's purpose-limitation rules do not apply to it.

Once information has been converted into an anonymised form, the data protection - and many other - restrictions on its use fall away. This provides an obvious incentive for organisations to produce and use anonymised information where this is a viable alternative to using personal data for a particular purpose.

There is clear legal authority for the view that, where a data controller converts personal data into an anonymised form and publishes it, this will not amount to a disclosure of personal data - even though the disclosing organisation still holds the 'key' that would allow re-identification to take place. This means that the DPA no longer applies to the disclosed information. This provides an obvious rationale for organisations that want to publish information to do so in an anonymised form - and for researchers and others to use anonymised information as an alternative to personal data wherever this is possible.

Personal data and identification

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

For consultation

- The concept of 'identify' – and therefore of 'anonymise' - is not straightforward because individuals can be identified in a number of different ways.
- You may be satisfied that the information your organisation intends to release does not, in itself, identify anyone. However, in some cases you may not know whether other information is available that may allow re-identification by a third party to take place.

It can be difficult in practice to determine whether information has been anonymised or is still personal data. This code describes ways of assessing and mitigating the risks that may arise, particularly in terms of assessing whether other information is available that is likely to facilitate re-identification. In some cases, it will be relatively easy to determine whether it is likely that a release of anonymised information will allow the identification of an individual. In other cases it will be much harder, but the decision still has to be made.

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

In February 2005, the ProLife Alliance made a Freedom of Information Act (FOIA) request to the Department of Health for detailed statistical information about abortions carried out in 2003.

The Department of Health refused the ProLife Alliance's request 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.

Mr Justice Cranston held that the fact that the Department of Health had access to all the information from which the statistical information was derived, did not prevent it from processing the data in such a way that it could become data from which a living individual could no longer be identified.

In converting the underlying information into statistics, the Department of Health had effectively anonymised the information so that, taking account of all the means reasonably likely to be used, anyone receiving the statistics would not be able to identify any of the individuals to whom the statistics related.

3 Ensuring the effectiveness of anonymisation

The availability of 'other' information

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.

27% of successful applicants for the role of store supervisor at Partridge's Clothiers Ltd were educated to at least degree level.

Whilst this information may be derived from an examination of appointees' personnel files – their personal data – this information does not identify or relate to any particular individual so is not in itself personal data.

However, the Data Protection Act (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. This means that when assessing whether information has been anonymised effectively it is necessary to consider whether other information is available that – in combination with the anonymised information – would result in a disclosure of personal data. This can be an issue for the initial organisation – which may well want to keep certain information in an anonymised form – and for third party organisations which may wish to avoid the problems and liabilities that can result from re-identification.

The 'other information' 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 known

For consultation

what data is already available or what data may be released in the future. It is also generally unfeasible to see data return (i.e. recalling data or removing it from a website) as a safeguard given the difficulty, or impossibility, of securing the deletion or removal of information 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.

Freedom of information and personal data

The DPA is primarily concerned with the risks associated with the identification of individuals by data controllers. The right to ask for any information under the Freedom of Information Act (FOIA) makes it harder for a public authority to control the nature of the disclosure. However, section 40 of 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 information to *a member of the public* would breach the data protection principles. This is intended to encourage public authorities to be cautious when disclosing information due to the likelihood of other data controllers or members of the public, combining information to produce information that does *relate to and identify* a particular individual and which is therefore personal data.

The test in FOIA can be 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 (see p18) can help to address this problem.

It is good practice to try to look at identification 'in the round', i.e. all organisations disclosing anonymised information should assess whether any organisation or member of the public could identify any individual from the information 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.

It is also good practice to review this periodically, bearing in mind that subsequent data releases and the development of new techniques may facilitate re-identification that was impossible previously. Where a data-linkage process results in re-identification, the organisation holding the 're-constituted' personal data will take

For consultation

on its own data protection obligations. This could include preventing the continued availability of the personal data – although this may not be feasible where the personal data has already been published.

Borderline cases

There will clearly be borderline cases where it will be difficult, or even impossible, to determine whether there is reasonable likelihood of re-identification taking place. The test in the DPA for determining whether information is 'personal data' is based entirely on the identification or likely identification of an 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 information is personal data.

However, in genuinely borderline cases, it is good practice to take a more cautious approach to disclosure where particularly sensitive information - or information that can significantly affect an individual's privacy - is involved. However, it is wrong to adopt the approach of categorising information that is sensitive, risky or consequential for individuals as personal data but other information as not personal data. Issues to do with the risk to individuals should normally be considered later when assessing whether disclosure would breach the data protection principles, not when making the initial assessment as to whether or not information is personal data.

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

- seek data subject consent for the disclosure of the information, explaining its possible consequences;
- adopt a more rigorous form of risk analysis and anonymisation; and
- in some scenarios, only disclose within a properly constituted closed community and with specific safeguards in place.

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 information is released, rather than the request being turned down entirely on

For consultation

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 where this would breach the data protection principles.

What is the risk of re-identification?

By 're-identification' we mean taking information that does not in itself identify anyone – 'anonymised information' - and analysing it or combining it with other information so that an individual is identified.

In some cases the risk of anonymised information being combined with other information to result in personal data being created will be high. An obvious example is where publicly available information – such as the Electoral Roll or information easily retrievable from a web-search – can be combined with the 'anonymised' information, 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 information 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 information derived from individuals' personal data. This can be a particular vulnerability where pseudonymised data sets are concerned, because even though pseudonymised information 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 information 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.

Data controllers must be aware of the risk of re-identification and that this risk can change over time. However, if anonymisation is carried out effectively in the present this is likely to protect personal

For consultation

data from future re-identification attack. A realistic assessment of the risk of re-identification occurring in the future should be made, meaning that organisations should not assume that information 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.

Publicly available information and anonymisation risk

An electoral roll entry includes the name, address and date of birth of those eligible to vote:

Mrs K L Thomas: 1 Sandwich Avenue, Stevenham, SV3 9LK. DoB 23/02/1960.

The public availability of the Electoral Roll 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.

The public availability of Mrs Thomas' date of birth might also present a re-identification threat, where, for example, an 'anonymised' research database is published that contains the complete dates of birth and partial postcodes of research subjects.

Customers' purchasing habits: linking anonymised information

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 that will involve correlating shoppers' purchasing habits with public health data about diabetes rates.

In order to do this Buysome and other local supermarkets use an encryption algorithm 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 keep the decryption key needed to link the unique reference number to a particular individual secure.

This results in two anonymised datasets that the researchers can match together and analyse even though they cannot identify any individual. In this case no personal data has been disclosed, even though both Buysome and the local GP surgeries retain the personal data from which the reference numbers were derived.

Re-identification risk

The University of Stevenham Research Centre (USRC) receives a request from a neighbouring research centre, doing similar research on the relationship between individuals' time on incapacity benefit, age range and body mass index (BMI). USRC decides to disclose the following anonymised extract from its dataset:

1. Name, address, date of birth	2. Period of Incapacity Benefit.	3. Body mass index	4. IB / BMI correlation score	5. Age range	6. Research cohort reference no.
[REDACTED]	< 2 years	20	[REDACTED]	40-45	1A5
	> 5 years	21		50-55	2B4
	< 2 years	22		40-45	3C3
	> 5 years	23		45-50	4D2
	< 2 years	24		45-50	5E1

It performs the anonymisation through a simple process of redaction – in this case deleting the content of certain columns. USRC could use the following key to re-identify *Figure 1* type information.

Name: Mr B Stevens
Address: 46 Sandwich Avenue, Stevenham, SV8 6PR
Research cohort ref. no. = 1A5

Any organisation with the information in *Figure 1* and access to this 'key' would clearly be able to discover that Mr B Stevens has been on incapacity benefit for less than 2 years and has a BMI of 20. However, the re-identification process is only possible here because the information in the table has a particular characteristic: it is divided into separate data fields that relate to particular individuals, allowing other information to be 'mapped on' to it resulting in re-identification. The 'mapping' process – whereby information is linked with a high degree of certainty to other information held in an anonymised form – is not possible where the anonymised information is no longer separated into fields that relate to a particular individual. This might be the case where aggregated information is produced from a set of personal data.

However, this is still a safe form of identification because only USRC holds the 'key', it is not in the public domain and, given proper organisational and technical security, USRC should be able to control access to it.

For consultation

The 'motivated intruder' test

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

- a) the anonymised information they release will allow the re-identification of the individuals; or
- b) whether anyone would be likely to do this in practice.

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

The 'motivated intruder' is taken to be a person who starts without any prior knowledge but who wishes to identify the individual or individuals referred to in the information to be released and who will take all reasonable steps to do so.

Application of the 'motivated intruder' test involves considering whether a 'motivated intruder' (as a representative of the public at large), would be able to identify the individual to whom the disclosed information relates.

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.

The 'motivated intruder' test assumes that, regardless of the apparent attractiveness or interest-worthiness of the information to be disclosed, there will be someone (a 'motivated intruder') who would want to identify the individuals to whom it relates and who will use all methods reasonably available to do so. Consequently, careful consideration should be given to the risks associated with the disclosure of even 'ordinary' or 'innocuous' information.

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

- financial gain or commercial advantage;

For consultation

- the possibility of causing mischief by embarrassing others;
- revealing newsworthy information about public figures;
- a hacker wanting to prove that a 'hack' of anonymised data is possible;
- political or activist purposes – for example as part of a campaign against a particular organisation or person; and
- curiosity – for example 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 information 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 a high level of risk to individuals should re-identification occur. One example might be health information, 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 information should reflect this. In reality, though, information 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 inexperienced' 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.

It is also 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 information 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.

Motivated intruder risk: some issues to consider

- What is the risk of jigsaw attack? Does the information have the characteristics needed to facilitate data linkage?
- 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 vulnerabilities did it reveal?

'Motivated defenders'

As explained above, it can be difficult to assess the risk of re-identification by a member of the public because different members of the public – and indeed different organisations – have access to different information resources, potentially much richer ones, than the general public. The 'motivated defender' test will be useful in certain circumstances and is intended to encourage organisations to consider how those with the specialist knowledge needed to perform re-identification might react in response to a disclosure of linkable information.

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 information 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; or
- 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 information 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 have to have access to so much information about the other individual for re-identification to take place. The situation is similar where an individual might recognise that anonymised information relates to him or her, allowing self-identification to take place.

It could be the case that those who have an unusual amount of knowledge about someone else will want to prevent the identity of the individual concerned being disclosed because of family loyalty or professional obligations, for example. Such individuals will be motivated to defend the identity of the individuals concerned. The term ‘motivated defender’ has been coined to describe those who have prior personal knowledge that would allow them to identify an individual but who, for personal or professional reasons, would defend the individual’s identity from others.

In other circumstances, those with prior personal knowledge which would allow them to identify the individual(s) to whom the anonymised information relates, may have either no reason to ‘defend’ the individuals’ identities or may wish to actively publish information about them. An individual, who finds out something about somebody else through a combination of anonymised information and prior personal knowledge, could decide to tell others or post it on the internet. Such individuals with prior personal knowledge are not ‘motivated defenders’.

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

- whether there are any particular individuals with the knowledge necessary to allow re-identification to take place (individuals with prior personal or professional knowledge);
- how likely it is that the anonymised information will come to their attention or be sought out;
- how someone capable of carrying out re-identification is likely to act; and
- what the consequences of re-identification are likely to be, if any, for the data subject concerned.

For consultation

It is generally reasonable to assume that those constrained by professional or legal obligations and close family and friends, and in some circumstances, colleagues, are likely to be motivated to defend the identity of an individual to whom the anonymised information relates. However, such assumptions must be checked in the light of any relevant additional information you may have. If, for example, you are advised that an individual has fallen out with his wife and that she would be only too keen to publish embarrassing personal data about him should she discover it, by being able to identify the anonymised information as relating to her husband, you should assume that the disclosure of the anonymised information would amount to the disclosure of personal data.

Information, established fact and knowledge

When considering re-identification risk, it is worth noting that the test in the DPA is whether it is reasonably likely that a combination of *data and other information* will allow an individual to be identified.

It is useful to draw a distinction between recorded information, established fact and personal knowledge. 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 present privacy risks.

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 regarding an individual's identity may present a privacy risk but not a data protection one. Even where a guess based on anonymised information 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 information 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 information should be withheld for some other reason, as discussed in section 6.

For consultation

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 information can present privacy risks even if it does not constitute personal data and cannot be converted back into personal data. This might be the case where the anonymised information points to a number of individuals, for example 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, 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 those individuals. 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 indicates that someone living in a small geographical area has 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 are 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.

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

The publication of personal data based on an individual's properly informed consent is unlikely to breach the data protection principles. However, there are obvious problems in this approach – particularly where an individual decides to withdraw consent. In reality, it may be impossible to remove the information from the public domain, so that the withdrawal of consent will have no effect. It is far 'safer' to publish anonymised information than personal data, even where consent could be obtained for the disclosure of personal data. It is worth noting that the 'necessity' rules in the Data Protection Act (DPA) mean that it could be against the law for an organisation to publish personal data where anonymised information could serve the same purpose.

Anonymising personal data constitutes 'processing' for the purposes of the DPA, and therefore needs to be justified by one or more of the DPA's conditions for processing. The DPA provides various bases for legitimising the processing of personal data – including its anonymisation. Consent is one of these. However, the DPA does not necessarily require consent for the creation of anonymised data or its disclosure.

Obtaining consent can be logistically very onerous – for example where large numbers of personal records are involved. It could even be impossible – for example where the personal data is old and there is no reliable means of contacting individual data subjects. In the Information Commissioner's view it is generally acceptable to anonymise personal data and to disclose it without the data subject's consent, provided that:

- the anonymisation will be done effectively, with due regard to any privacy risk posed to individuals – a privacy impact assessment² can be used here;
- the purpose for which the anonymisation takes place is legitimate and has received any necessary ethical approval;

²http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/privacy_impact_assessment.aspx

For consultation

- neither the anonymisation process - nor the use of the anonymised information - will have any direct detrimental effect on any particular individual;
- the data controller's privacy policy – or some other form of notification - explains the anonymisation process and its consequences for individuals; and
- there is a system for taking individuals' objections to the anonymisation process or to the release of their anonymised information into account. [Note though that the DPA does not give individuals a general right to prevent the processing (including the anonymisation) of information about them. It is good practice though to respect individuals' objections where possible, and may be a requirement of the DPA where convincing reasons are present.]

5 Personal data and spatial information

There is no simple rule for handling postcodes and other geographical information under the Data Protection Act (DPA); in some circumstances this will constitute personal data – for example 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.

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

In some cases it may be necessary to process postcodes, removing certain of their elements, to reduce the risk of identification. 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

('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'.)

Source: Centre for Advanced Spatial Analysis: UCL

With information relating to a particular geographical area, there can be a distinction between a “statistical comfort zone” that eliminates 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 FOIA request a different approach may be possible, based on an application of the tests in the DPA.

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 spatial datasets to the public.

Some principles for the publication of spatial information

- The larger the number of properties or occupants in a mapping area, the lower the privacy risk.
- Privacy risk depends on the frequency of publishing data and the way it is represented and categorised. Where real-time or very frequent publication takes place, it becomes easier to link geographical data to an individual. Publishing data very frequently or in real-time poses a greater privacy risk.
- Publishing geographical information on a household level could constitute the processing of personal data and, depending on the circumstances, could breach the first data protection principle’s requirement of fairness. This is because it is quite easy to link a property to its occupant or occupants – using the publicly available Electoral Roll, for example.
- The use of heat maps, blocks and zones can help to present valuable geographical information in a way which reduces privacy risk. A strong public interest case would have to be made for the use of more granular or intrusive indicators.
- For public authorities transparency should be the default position regarding the publication of geographical information. 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

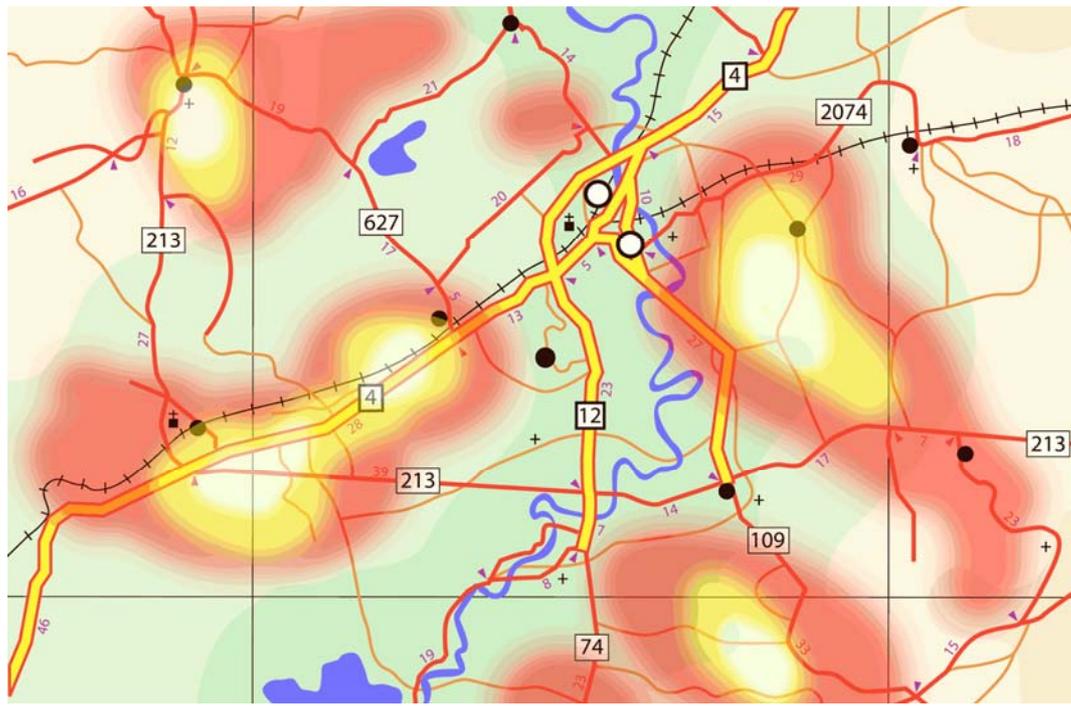
For consultation

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

Compliance with the Data Protection Act (DPA) should not be the only consideration when considering whether information can be made available. In some scenarios data controllers (often public authorities) may make a decision that the information they have been asked to disclose is not personal data. However, this does not necessarily mean that the information in question should or must be disclosed. Disclosure of the information may still present a risk to individuals. For example, a risk may arise where supposition or 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. This could be a real factor where 'high risk' data is involved.

Information that identifies an organisation but not an individual is generally not personal data. However, it may still be legitimate to withhold information identifying an organisation on the grounds of commercial confidentiality.

It is important not to attempt to extend the definition of personal data to cover scenarios where, in reality, no information that relates to an identifiable individual is involved. Another exemption may prevent the disclosure, and you should remember that if information is personal data for the purposes of the Freedom of Information Act (FOIA), then it is also personal data for the purposes of the DPA. This means that all the DPA's provisions – for example, subject access rights and 'fair processing' rules – will apply to the information.

Human rights

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

For consultation

The Article 8 right will often overlap with the protection provided for by the DPA; if the 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 engage the broader provision in the HRA. For example, disclosing information about a large family group might not disclose personal data but may well breach the privacy rights of the family. It is advisable to seek specialist advice if you believe a disclosure has Article 8 implications.

Freedom of Information Act (FOIA) exemptions

In some cases you may still need to consider the risk to health and safety of members of the public, even if you have decided the information to be disclosed is not personal data. Under FOIA this may require consideration of the section 38 'health and safety' exemption. The Commissioner has accepted this approach in a FOIA case involving the disclosure of information about a group of children in care.

In some scenarios the Commissioner accepts that the impact of disclosure, whilst not leading to identification, may still change behaviour (e.g. participation in a survey) through fear of educated guesses or supposition. If the impact of this falls on public authorities' function or core activities it will be reasonable for them to consider FOIA's section 36 'prejudice to effective conduct of public affairs' exemption.

Other statutory prohibitions

Other statutory prohibitions may also 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 information. Under FOIA a breach of a statutory prohibition would engage the section 44 exemption.

Statistical confidentiality

Compliance with the Statistics and Registration Service Act 2007 and following Code of Practice for Official Statistics must be considered by central government departments when disclosing statistics.

7

Publication and limited disclosure

It is not always necessary, or desirable, to publish even anonymised information. Clearly the open data agenda relies on the public availability of information, and information released in response to a Freedom of Information Act (FOIA) request cannot be restricted to a particular person or group. However, much research, in particular, takes place through releasing data within a closed community. The obvious advantage of this is that re-identification and other risk is more controllable, and therefore more information can be disclosed without having to deal with the problems that publication can cause.

Means of making information, whether anonymised or not, available to third parties or the general public include the following.

- **Publication.** This is where information is made publicly available and anyone can see it and, in reality, use it for their own purposes. This can further transparency and deliver other benefits but once published no strict controls can be placed on re-identification, although other elements of the law may still apply - for example where information is subject to copyright. However, any third party performing re-identification will take on its own data protection liabilities. In reality, publication under licences such as the Open Government Licence falls into this category, as do disclosures made under FOIA or the transparency agenda. (The Open Government Licence does not apply to the use or reuse any personal information contained in a publication.)
- **Publication under specific licence terms.** This is an attempt to make information publicly available but to place certain specific restrictions on the way it is used. Whilst this can provide useful protection in respect of recipients that respect the rules, this form of publication can clearly present a privacy risk if the conditions attached to the information are either unlikely to be respected or not enforceable.
- **Access control.** This is where anonymised information – or in some cases personal data – is disclosed but only to particular recipients, with conditions attached to the disclosure. This is often used between groups of researchers. It is appropriate for handling anonymised information that is particularly sensitive in nature or where there is a significant risk of re-identification. The great advantage of this approach is that the disclosing

For consultation

organisation can impose rules – in some cases legally binding ones – and practical restrictions relating to such matters as security and purpose limitation, and can prohibit any attempt at re-identification. (This approach is not viable for FOIA requests.)

There will be cases where a piece of research or planning, for example, depends on the researchers having access to information that cannot be aggregated, blurred or reduced to statistics. It may be the case that the information is particularly high-risk and that it is not clear whether it can be anonymised effectively. In other cases access to personal data may be necessary - although it is difficult to envisage situations where explicit identifiers such as names and addresses cannot be replaced with less obvious ones. The risks in situations like this can be reduced by only making the information available within a 'closed' community – for example, to a research body. This allows the organisation disclosing the information to place conditions on its use and to put safeguards in place.

Safeguards should include:

- limitation of the use of the information to a particular project or projects;
- restriction on the disclosure of the information;
- prohibition on any attempt at re-identification;
- arrangements for technical and organisational security, for example 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.

Safety versus utility

The key to performing an anonymisation process effectively is to ensure that the resultant anonymised information does not have the properties, or vulnerabilities, that can facilitate data linkage and lead to re-identification. It is good practice to use perturbation techniques to 'blur' information so that it cannot be matched reliably with other data regardless of how sensitive or innocuous the data is. Many types of aggregated releases will be relatively low-risk, depending on granularity, sample sizes and so forth. However, there is a balance to be struck between the utility of data, for example to planners or researchers, and its protection from re-identification. The methods used to protect data can be very effective, but are often difficult to apply. In order to have good reason to believe that the correct method has been used with the correct settings, it may be necessary to obtain expert advice. Even for the most sophisticated methods, it remains a general rule that the better protected the information is, the lower its utility. (Although certain statistical information with low or no risk of re-identification, can be sufficient for certain purposes and pseudonymisation can provide significant privacy protection without necessarily lowering its utility for a longitudinal study.) This is why disclosure within a closed community, ideally using pseudonymised data, rather than publication, may be the appropriate option where individual-level, match-able data is needed – for example, to carry out a longitudinal study of individuals' progress through the education system and into employment.

Appendix 1 contains some useful worked examples of anonymisation but it goes beyond the scope of this code to assess the utility of information that has been anonymised using particular methods to researchers and others. However, it is worth noting that information that has been subjected to perturbation methods, for example, can alter the structure and detail of the data, reducing its usefulness. It is also important to recognise that anonymisation techniques should not be applied on a 'one size fits all' basis. If they are used, it needs to be in discussion with the data recipient to tailor the data to simultaneously comply with relevant requirements and to meet the needs of the project for which the data is to be supplied.

8 Governance

If your organisation is involved in the anonymisation and publication of information, it is good practice to have a governance structure in place that will address the whole process of producing and publishing anonymised information.

Having a suitable governance structure in place will be relevant should the Information Commissioner's Office (ICO) receive a complaint about your processing of personal data, including its anonymisation. 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 information it published did not contain personal data or present a re-identification risk.

A governance structure should cover the following points.

- 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 and to ensure co-ordination between technical, legal and policy staff. A 'Senior Information Risk Officer' (SIRO) approach can be particularly useful. It is important to train staff so they have a clear understanding of anonymisation techniques, any risks involved and the means of mitigating these.
- What is your procedure 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.
- How do you assess privacy impact? Many organisations involved in the creation or disclosure of anonymised data will find the ICO's Privacy impact assessment (PIA) handbook a useful way to structure their decision-making process. The approach in the handbook can easily be read across to cover many anonymisation scenarios. The Information Commissioner

recommends that, where possible, organisations publish their PIA report to show the public how they have approached the risk-assessment process.

[Read the ICO PIA handbook³](#).

- How are you going to explain your anonymisation of personal data to data subjects? Whilst it may not be necessary, and may even be impossible, to contact individual data subjects, your organisation's privacy policy – which should be clear and easily accessible to the public - should explain your organisation's approach to anonymisation and any consequences of this.
- What form of disclosure do you intend to use? Do you intend to publish anonymised information so that it is publicly available or would a more limited form of disclosure be appropriate?
- How are you going to review the consequences of your anonymisation programme, particularly any privacy risk to individuals? 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 also 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.

Re-identification testing

It is good practice to use re-identification testing – a type of 'pen' or 'penetration' testing - to detect and deal with re-identification vulnerabilities related to the creation or publication of anonymised information. Pen-testing, i.e. attempting to re-identify individuals from the relevant anonymised information, can be particularly useful where an organisation is finding it difficult to assess the risk posed to individuals by re-identification.

There can be advantages in using a third party organisation to carry out the testing, as it may be aware of information resources, techniques or types of vulnerability that you have overlooked or are not aware of. The further reading section below details an example of pen-testing conducted by University of Southampton for the Ministry of Justice on reoffending data.

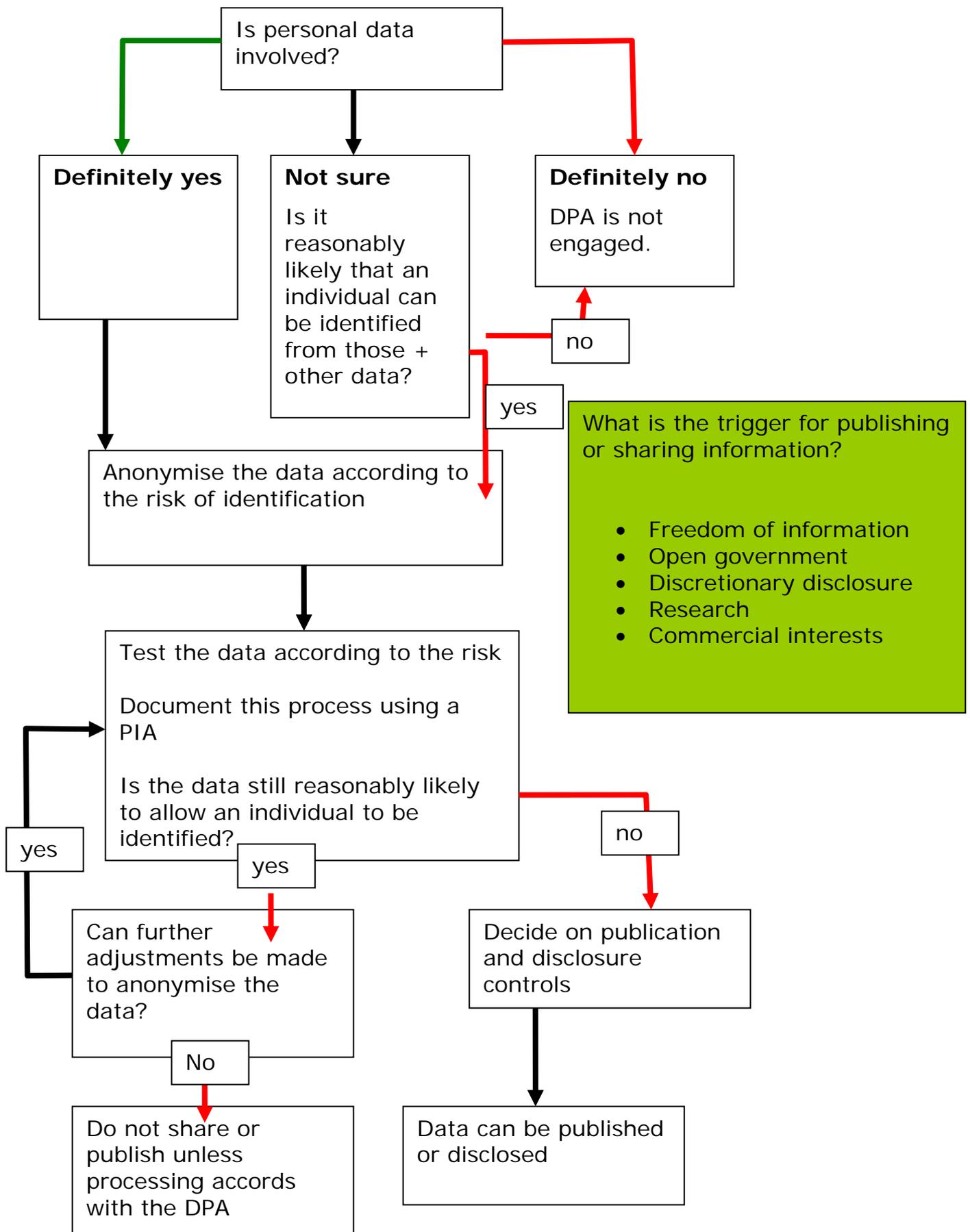
3

http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/privacy_impact_assessment.aspx

For consultation

The first stage of a re-identification testing process is to take stock of the anonymised information that your organisation has published or intends to publish. The next stage is to try to determine what other information, personal data or not, is available that could be linked to the anonymised information to result in re-identification. As we have explained elsewhere in this code, 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 have access to. However, you can 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 above can also form a useful component of a pen-test.

The situation becomes more complex where statistical information is involved, because there may be various statistical data sets publicly available which, if matched in a particular way, could result in re-identification. 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. Note, though, that even if re-identification testing shows that it is possible to determine that one set of anonymised data, for example in statistical form, relates to the same individual as another set of data, this does not mean that identification has taken place or, therefore, that personal data has been created. However, the ability to determine that one set of data relates to the same individual as another is certainly a factor that can make re-identification more likely.



9 The Data Protection Act research exemption

There is an exemption from various provisions of the Data Protection Act (DPA) for personal data processed only for research purposes, provided certain conditions are satisfied. The DPA does not define 'research'. However, the Information Commissioner believes that to benefit from the exemption, the research in question must be a properly defined search or investigation undertaken to discover facts and increase knowledge. The DPA makes it clear that 'research purposes' include statistical or historical research, but other forms of research, for example social research, could benefit from the exemption.

The conditions that must be satisfied are that:

- the data are not processed to support measures or decisions with respect to particular individuals; and
- 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.

Provided the data are 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;
- the data may be retained indefinitely; and
- the data will be exempt from the right of subject access – provided there is no publication of data in a form which identifies any individual or individuals.

Clearly the research exemption provides important benefits for researchers and important safeguards for individuals. However, it is still good practice to anonymise personal data as early in the research process as possible – ideally before it is disclosed or used for research purposes. This minimises, or negates, the risk to individuals and means that researchers will not need to be concerned with the parts of the DPA from which section 33 does not provide exemption.

The section 33 exemption can still be used even if research outputs are published in a form which identifies individuals. However, depending on the nature of the data, this could still breach other provisions of the DPA – for example the first data protection principle's requirement of fairness and lawfulness in the processing of personal data and the need to satisfy a 'condition for processing'. It is certainly good practice to avoid the publication of research data in a form which identifies individuals where there are alternatives to this, as there generally will be.

There is a particular incentive to anonymise sensitive personal data – for example information 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 subject consent for the anonymisation of personal data about them, but the DPA does provide 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 bad practice – and arguably a breach of the DPA – to publish or disclose 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 information.

Appendix 1:

Practical examples of some anonymisation techniques – drawn up by Mu Yang, Vladimiro Sassone and Kieron O’Hara at the University of Southampton.

Data reduction

1. Removing variables

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

Description: The simplest method is the removal of 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 too sensitive for public use or irrelevant for analytical purpose.

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.

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

For consultation

2. Removing records

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	F	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	F	SO14	£28,000	£1,200

Income & Expenses individual-level dataset				
Age	Sex	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	F	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	F	SO14	£28,000	£1,200

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.

Comments: Removing records is similarly damaging to the information content of the matrix to removing variables; the former removes a column from the table, while the latter 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 (i.e. distort) the statistical properties of the released data.

3. Global recoding

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

Description: This method makes variable values less specific, and the table therefore less informative. For a categorical variable, 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 is a discrete version of the original 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, very high or very 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.

Example: In example three the variables 'Age' and 'Postcode' are aggregated into new classes, each of which has values as a range. Each more specific value has a unique mapping 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.

For consultation

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).

4. Local suppression

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	SO17-19	low	low
25-29	M	SO17-19	low	low
25-29	M	SO14-16	medium	medium
40-44	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
40-44	F	SO14-16	medium	low

The combination "Age =20-24, Gender =F" is 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)
missing	F	SO17-19	low	low
25-29	M	SO17-19	low	low
25-29	M	SO14-16	medium	medium
40-44	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
40-44	F	SO14-16	medium	low

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. When combinations of such variables are problematic, local suppression consists of replacing an observed value with a missing value. 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.

For consultation

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 suppress the variable ‘Age’ of this record as ‘missing’.

Comments: Local suppression should be applied only to risky records (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.

Data perturbation

5. Micro-aggregation

Micro-aggregation

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	SO14	£28,000	£1,200
35	F	SO17	£31,500	£2,000
30	M	SO16	£32,000	£1,800
40	F	SO15	£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	SO14	£23,333	£1,200
35	F	SO17	£43,833	£2,000
30	M	SO16	£43,833	£1,800
40	F	SO15	£43,833	£3,500

k -partition = 3

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

For consultation

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 with n records, these records are combined (usually in their size order) to form g groups which have 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.

Micro-aggregation can be independently applied to one variable or a set of variables. It is then called *individual ranking*. 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 *3-partition* (i.e. we set k to 3). So the group number g in this small example of 6 individuals is 2. Then we compute the average value for each group and replace the original value by the average value.

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

6. Data swapping

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	Gender	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
25-29	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
40-44	F	SO14-16	medium	low

value unique

Income & Expenses individual-level dataset				
Age	Gender	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
40-44	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 because of the limited number of records.

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%. In some situations there may be conditions on the swapped pairs of records, which are constraints on the variables which are not swapping attributes, 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*. 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

For consultation

rate $r = 2/6 = 33.3\%$. The high value of the swapping rate is of course due to the small population in the example.

Comments: Swapping does not change the distribution of any variable, but still there is a tradeoff – lower risk implies higher information loss.

7. Post-Randomisation Method (PRAM)

Post-Randomization 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: $p_{11}=p_{22}=0.9$, $p_{12}=p_{21}=0.1$.

Description: The Post-Randomization Method is a probabilistic, perturbative method for categorical variables in data files. 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 complex probability mechanism which is named a *Markov matrix*.

The basic principle is as follows. Suppose we have a categorical variable V , which we wish to perturb. 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 $K \times K$ matrix which is the Markov Matrix. The individual entries in the Markov Matrix are referred to as p_{11} , p_{12} , p_{13} , p_{21} , p_{31} , etc, so that, say, p_{31} is the probability that category 3 of variable V will be transformed into category 1 of variable X .

For consultation

in the perturbed file. The general case, the probability of transforming k into l is referred to as p_{kl} .

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 $p_{k1} \dots p_{kK}$. 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 $p_{11} = p_{22} = 0.9$ 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, two of these three 'males' were originally male and similarly, two of these 'females' were originally male.

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.

For consultation

8. Adding noise

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



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

Description: Adding noise, a method applied to numerical data, consists in adding a random value ϵ , with mean zero and predefined variance σ^2 , to all values in the variable to be protected.

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.

9. Resampling

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

grand mean: £33,208

Normal distributed resampling,
Hypothesis testing: Null Hypothesis.

Income individual-level dataset				
Age	Sex	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

grand mean: £33,208

Income individual-level dataset with ordered mapping				
Age	Sex	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

Description: Resampling is also designed for numerical data. Broadly speaking, it has three steps. First, identify the probability density function of the sensitive data variables across the whole population, not just the database (e.g. is it a normal distribution, a Poisson distribution, etc?), and estimate the associated parameters (e.g. mean, variance) of the density. Second, generate a distorted sample with these parameter values. Third, replace the confidential data with the distorted sample. The sample should therefore be the same size as 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 cumulative distribution 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

For consultation

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 (e.g. the correlation between age and income) are preserved to some extent.

Comments: In reality, estimating the probability density function of the variables in the original data file would be impossible to achieve with complete accuracy, as sufficient information about the true distribution of data is not available. The data will only sometimes follow a specific theoretical distribution, which may make creating the distorted sample more difficult. Information about individuals is lost, and the correlations between variables will be affected.

Reference: In this survey, the authors are indebted to the account in Molla Hunegnaw, African Centre for Statistics, [*Confidentiality and Anonymization of Microdata*](#).

Appendix 2: the data protection principles

1. Personal data shall be processed fairly and lawfully and, in particular, shall not be processed unless;
 - (a) at least one of the conditions in Schedule 2 is met, and
 - (b) in the case of sensitive personal data, at least one of the conditions in Schedule 3 is also met.
2. Personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.
3. Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed.
4. Personal data shall be accurate and, where necessary, kept up to date.
5. Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.
6. Personal data shall be processed in accordance with the rights of data subjects under this Act.
7. Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.
8. Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data.

Appendix 3: Anonymisation techniques

Technique	Description
De-identification	<p>This involves stripping out formal personal identifiers – such as names - from a piece of information, to create a data set in which no person identifiers are present.</p> <p>Variants:</p> <ul style="list-style-type: none"> • Partial de-identification – results in information where some personal identifiers e.g. 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 information needed to facilitate re-identification. It can involve disclosing unique personal identifiers – e.g. reference numbers – but not the ‘key’ needed to link these to particular individuals.
Pseudonymisation	<p>De-identifying information so that a coded reference or pseudonym is attached to a record to allow the information to be associated with a particular individual without the individual being otherwise identified.</p>
Aggregation	<p>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.</p> <p>Variants:</p> <ul style="list-style-type: none"> • Inference Control - Small cell values (e.g. 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.

	<ul style="list-style-type: none">• 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.• 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 are not related to any particular individual.
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 e.g. 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.

Appendix 4: Glossary

Aggregated information

Statistical information about multiple 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.

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 information from a variety of sources, typically information that relates to the same individual.

Data subject

An individual who is the subject of personal data.

Disclosure

The act of making information or 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.

Longitudinal study

For consultation

A study that involves linking data about the same individual over a period of time – for example to study an individual's health episodes.

Open Data

Datasets that are made accessible in non-proprietary formats under licenses that permit unrestricted re-use (OKF – Open Knowledge Foundation, 2006)

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 information or data publicly available.

Re-identification

The process of analysing data or combining it with other data with the result that individuals become identifiable.

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,

For consultation

- (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 information

Information which is held in the form of numerical data, which may or may not allow individuals to be identified.

Appendix 5: 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](#)
- [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.
- [Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modelling Performance](#)
- [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' 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)

For consultation

- [Introduction to Privacy-Preserving Data Publishing Concepts and Techniques](#)
- [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](#)