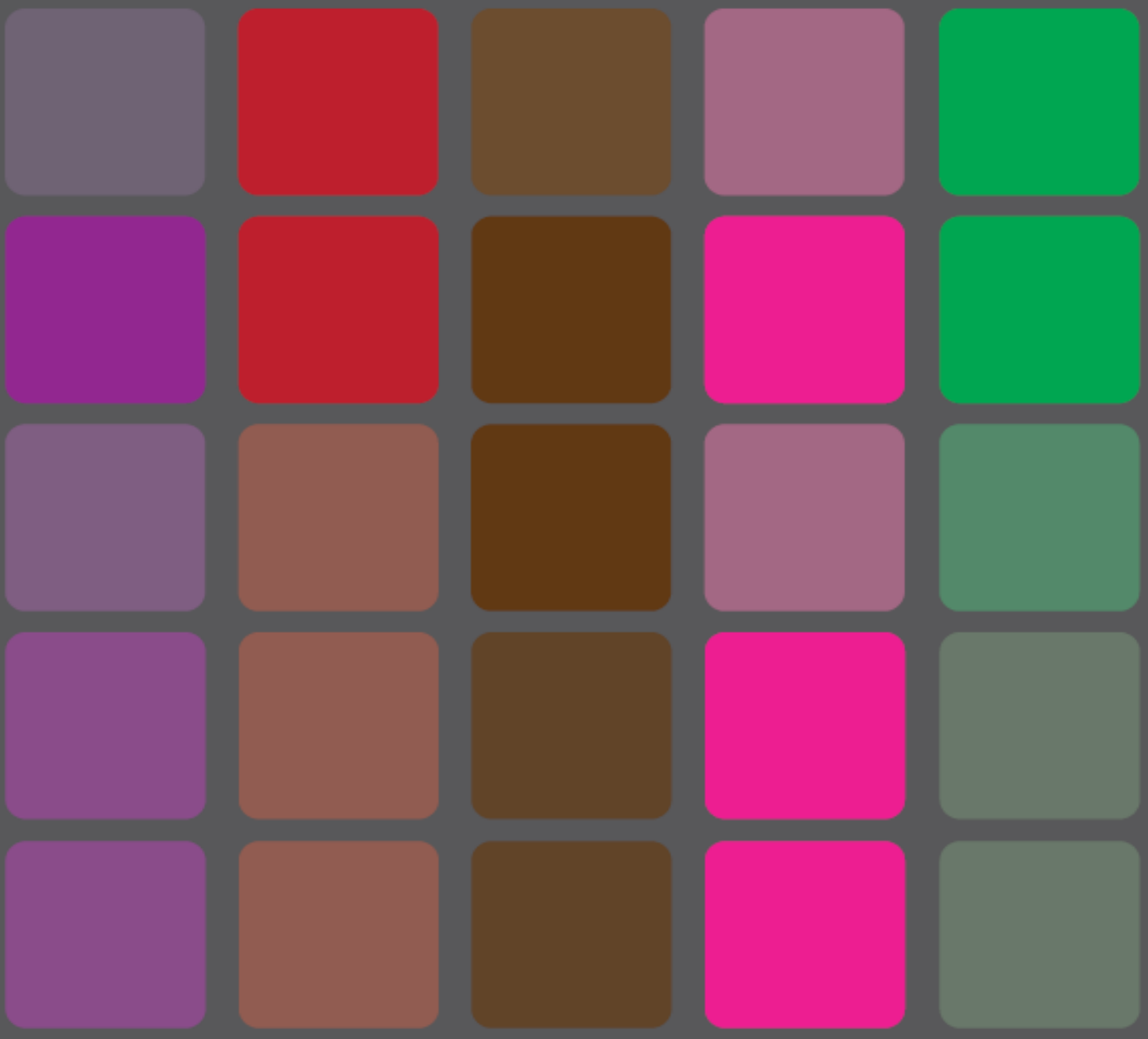
THE ANONYMISATION DECISION-MAKING FRAMEWORK

Mark Elliot, Elaine Mackey
Kieron O'Hara and Caroline Tudor



UKAN PUBLICATIONS

The Anonymisation Decision-Making Framework

Mark Elliot, Elaine Mackey
Kieron O'Hara and
Caroline Tudor

Published in the UK in 2016 by
UKAN
University of Manchester
Oxford Road
Manchester
M13 9PL

Acknowledgements

Part of the development work for the anonymisation decision making framework was conducted through a series of one-day meetings of members of the UKAN core network between 2012 and 2014 and we would like thank the group for their enthusiastic involvement:

- Ulrich Atz (Open Data Institute)
- Iain Bourne (Information Commissioner's Office)
- Nigel Dodd (Telefonica)
- Keith Dugmore (Demographic Decisions Ltd)
- Dawn Foster (NHS Information Centre)
- Paul Jackson (Office for National Statistics)
- Jane Kaye (University of Oxford)
- Fred Piper (Royal Holloway College, University of London)
- Barry Ryan (Market Research Society)
- Claire Sanderson (NHS Information Centre)
- Chris Skinner (London School of Economics)
- Nigel Shadbolt (Open Data Institute)
- Natalie Shlomo (University of Manchester)
- Sam Smith (Med Confidential)
- Peter Stephens (IMS Health)
- Linda Stewart (National Archives)
- Nicky Tarry (Department of Work and Pensions)
- Jeni Tennison (Open Data Institute)
- Steve Wood (Information Commissioner's Office)
- Matthew Woollard (University of Essex)

We would also like to thank the following people for their extensive and thoughtful feedback on a draft of this book:

- Iain Bourne (Information Commissioners Office, UK)
- Martin Bobrow (Wellcome Trust, UK)
- Paul Burton (University of Bristol, UK)
- Josep Domingo-Ferrer (Universitat Rovira i Virgili, Spain)
- Jörg Drechsler (Institut für Arbeitsmarkt und Berufsforschung, Germany)
- George Duncan (Carnegie-Mellon University, US)
- Khaled El-Emam (University of Ottawa and CEO of Privacy Analytics, Canada)
- Jon Fistein (Medical Research Council, UK)
- Christine O'Keefe (Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia)
- Malcolm Oswald (HSISC, UK)
- Natalie Shlomo (University of Manchester, UK)
- Jeni Tennison (Open Data Institute, UK)
- Mathew Woollard (University of Essex, UK)

Foreword

Anonymisation is a subject that, on the face of it, is very technical and rather academic. However, it is of great – and growing - real-world significance. My office deals every day with the tension between access to information and personal privacy. Done effectively, anonymisation can help us to manage that tension.

The ICO published its 'Anonymisation: Managing Data Protection Risk' code of practice in 2012. The work we did on that taught us two main lessons. Firstly, effective anonymisation is possible but it is also possible to do anonymisation ineffectively. Secondly, it isn't always possible to draw the definitive personal / non-personal data distinction that legal certainty in the field of data protection depends on.

We have learnt that we have to deploy effective anonymisation techniques and assess re-identification risk in context, recognising that there is a wide spectrum of personal identifiability and that different forms of identifier pose different privacy risks. This authoritative and accessible decision-making framework will help the information professional to anonymise personal data effectively. The framework forms an excellent companion piece to the ICO's code of practice.

It is easy to say that anonymisation is impossible and that re-identification can always take place. It is just as easy to be complacent about the privacy risk posed by the availability of anonymised data. It is more difficult to evaluate risk realistically and in the round and to strike a publicly acceptable balance between access to information and personal privacy. The guidance in this framework will help information professionals to do that.

The General Data Protection Regulation will be fully implemented across the EU in 2018, following protracted and complex debate about – amongst other things – the nature of anonymisation and the status of pseudonymous data. The approach taken in the ICO Code of Practice and this framework is fully consistent with the GDPR. We must conclude that some individual-level data is personal data but some is not, depending on factors such as the nature of the data and the difficulty or cost of rendering it identifiable. Again, this may not make for the absolute legal certainty many would like, but it does provide for the flexibility we need to make sensible decisions based on the circumstances of each case.

Given the development of increasingly powerful data sharing, matching and mining techniques – and a backdrop of strong political and commercial pressure to make more data available - it can seem inevitable that re-identification risk will increase exponentially. However, this framework demonstrates that the science of privacy enhancement and our understanding of privacy risk are also developing apace.

It is essential that we continue to develop anonymisation and other privacy enhancing techniques as an antidote to the potential excesses of the big data era. The ICO has been one of the strongest champions of the privacy enhancement agenda, within the EU and beyond, and it will continue to be. Our work with the UK Anonymisation Network and our support for the development of the Anonymisation Decision-Making Framework is indicative of the continued importance of this field of informatics to the fulfilment of the ICO's mission.

Elizabeth Denham
UK Information Commissioner

Contents

Preface

The need for well-thought-out anonymisation has never been more acute. The drive to share data has led to some ill-conceived, poorly-anonymised data publications including the Netflix,[1] AOL[2] and New York taxi[3] cases, underlining how important it is to carry out anonymisation properly and what can happen if you do not.

This book has been developed to address a need for a practical guide to anonymisation that gives more operational advice than the ICO's *Anonymisation Code of Practice*, whilst being less technical and forbidding than the statistics and computer science literature. The book may be of interest to an anonymisation specialist who would appreciate a fresh, integrated perspective on the topic. However, it is primarily intended for those who have data that they need to anonymise with confidence, usually in order to share it. Our aim is that you should finish the book with a practical understanding of anonymisation and an idea about how to utilise it to advance your business or organisational goals. To make this tractable we have focused on personal data and specifically on information presented in the form of a file or database[4] of individual level records; we have – for this edition at least – set aside the specialist topic of data about businesses.[5]

We present in full here for the first time *the Anonymisation Decision-Making Framework*, which can be applied, perhaps with minor modifications to the detail, to just about any data where confidentiality is an issue but sharing is valuable. However, the biggest demand for the framework is primarily from people and organisations dealing with personal data and so that is the focus of our exposition here.

We assume the regulatory context of current (2016) UK law, and you should bear in mind that other legal jurisdictions will impose different constraints on what you can and cannot do with data. Across jurisdictions there are differences in the

[1] See CNN Money (2010) http://tinyurl.com/CNN-BREACHES

[2] See Arrington (2006) http://tinyurl.com/AOL-SEARCH-BREACH

[3] See Atokar (2014) http://tinyurl.com/NYC-TAXI-BREACH

[4] Thus, at this present time we do not consider unstructured data. However, the principles of the ADF do apply to this type of data and we envisage widening the scope of the book to incorporate an examination of it in future editions.

[5] Business data has different technical properties and different legislation can apply to it.

interpretation of data protection legislation and in the meaning of key terms such as 'personal data'. A reader outside the UK context should interpret the legal-facing sections of the book with this in mind. That said, the fundamental premise of anonymisation, that it is designed to control the risk of unintended re-identification and disclosure, will hold regardless of the legal context and therefore the *principles* that the framework provides should be universally applicable. We have in places included pointers to other jurisdictions where we could do that without making the text cumbersome and future editions may attend to this issue in a more thorough manner.

The framework has been a long time in gestation. Its foundations are a twenty-year programme of research carried out at the University of Manchester, and the long-standing relationship between the University of Manchester and the UK Office for National Statistics (ONS). More recently, the authors of this book have been partners in the UK Anonymisation Network (UKAN)[6] which has driven forward the development of the framework and convinced us of the enormous demand for a book in this space. One aim of UKAN, and indeed this book, has been to integrate the many different perspectives on the topic of anonymisation and in particular to join up the legal and the technical perspectives. We would like to express gratitude to the contribution of the UK Information Commissioner's Office (ICO), which provided the seed funding for the network and has been actively engaged with its development.

Our view has always been that anonymisation is a heavily context-dependent process and only by considering the data and its environment as a total system (which we call the *data situation*), can one come to a well informed decision about whether and what anonymisation is needed. Good technique is important but without a full understanding of the context, the application of complex disclosure control techniques can be a little like installing sophisticated flood defences in the Atacama desert or, at the other end of the scale, not realising that building a house on the edge of a cliff is just a bad idea regardless of how well designed it is. Accepting the importance of context, it is also important to understand that a fully formed anonymisation process includes consideration of the ethics of data sharing and the importance of transparency and public engagement and you will find as you work through the book that the framework incorporates these elements too.

---

[6] UKAN provides services including training workshops and clinics for those who need to anonymise their data. These services can be accessed via the network website: www.ukanon.net.

We have decided to release this book as a freely available open source book rather than through a traditional publisher as we feel that we have an important message that we wanted to ensure is disseminated as widely as possible. We hope that you find the book of value. We would welcome comments on the book at any time via our web site www.ukanon.net. The book is intended to be organic and we will be updating it periodically.

Chapter 1: Introduction

In this chapter we introduce the Anonymisation Decision-making Framework (ADF), explaining the thinking behind it and the principles on which it is founded. We outline how you might best use the ADF (given your skills and experience) in your anonymisation practice. But first, let us make explicit the three central terms featured in this book: anonymisation, risk and sensitivity.

1.1 Anonymisation, risk and sensitivity

A common error when thinking about *anonymisation* is to focus on a fixed end state of the data. This is a problem because it leads to much muddled thinking about what it means to produce 'anonymised data'. Firstly, it exclusively focuses on the properties of the data whereas in reality whether data are anonymised or not is a function of both the data and the data environment. Secondly, it leads one into some odd discussions about the relationship between anonymisation and its companion concept risk, with some commentators erroneously (or optimistically) assuming that 'anonymised' means that there is zero risk of an individual being re-identified within a dataset. Thirdly, viewing it as an end state means that one might assume that one's work is done which in turn promotes a counterproductive mentality of 'release-and-forget'.

In some ways, it would be better to drop the adjectival form 'anonymised' altogether and perhaps talk instead of 'data that has been through an anonymisation process'. However, the constraints of the English language mean that this would sometimes lead to some quite tortuous sentences. So, in this book, we will use the term 'anonymised' but this should be understood in the spirit of the term 'reinforced' within 'reinforced concrete'. We do not expect reinforced concrete to be indestructible, but we do expect that a structure made out of the stuff will have a negligible risk of collapsing.

This brings us in turn to the notion of *risk*. Since Amos Tversky and Daniel Kahneman's seminal work in the 1970s, it has been clear that humans are quite poor at making judgements about risk and are subject to numerous biases when making decisions in the face of uncertainty (see for example Tversky and Kahneman (1974)). One aspect of this is the tendency to confuse the likelihood of an event with its impact (or disutility). To complicate matters further, where risks are dependent on human action, these biases themselves factor into the risk profile. So if we can

convince a data intruder that likelihood of a re-identification attempt succeeding is negligible then they are less likely to put the necessary effort in to attempt it and thus we have controlled the risk beyond what we have measured 'objectively'.

Thinking about the impact side of risk brings us to the third key concept, *sensitivity*, which tends to be connected with the potential harm of any confidentiality breach. However, as we will see, sensitivity is a larger concept than this and encompasses how the data were collected and what reasonable expectations a data subject might hold about what will happen to data about them.

Anonymisation, then, is a process of risk management but it is also a decision making process: should we release this data or not and if so in what form? Considering all the elements involved, that decision can appear complex and rife with uncertainties. It does require thinking about a range of heterogeneous issues from ethical and legal obligations to technical data questions: bringing all these disparate elements into a single comprehensible framework is what this book is all about.

1.2 The principles behind the ADF

The ADF incorporates two frames of action: one technical, the other contextual. The technical element of the framework will enable you to think about both the quantification of re-identification risk and how to manage it. The contextual element will enable you to think about and address those factors that affect re-identification risk. These include the particulars of your data situation such as the data flow, legal and ethical responsibilities and governance practices, your responsibilities once you have shared or released data, and your plans if, in the rare event, things go wrong.

The framework is underpinned by a relatively new way of thinking about the re-identification problem which posits that you must look at both the data and the data environment to ascertain realistic measures of risk. This is called the data situation approach. Perhaps it seems obvious that the environment in which data are to be shared and released is important, but for many years the data confidentiality field has focused almost exclusively on the data themselves. Thus re-identification risk was seen as originating from, and largely contained within, the data. As a consequence, researchers and practitioners rarely looked beyond the statistical properties of the data in question. With a few notable exceptions (e.g. Duncan & Lambert 1989, Elliot and Dale 1999 and Reiter 2005) they have not concerned themselves with issues such as how or why a re-identification might happen, or

what skills, knowledge, or other data a person would require to ensure his or her attempt was a success. As a consequence, the statistical models they built to assess re-identification risk, whilst statistically sophisticated, have at best been based on assumptions about the data context[7] and at worst totally detached from any real-world considerations.

To address these failings there have been attempts to describe and theorise about context beyond the data. This has usually taken the form of intruder scenario analysis which we will consider in more detail later in chapter 2 in 2.3.1 and in chapter 3 component 6 of the ADF. Scenario analysis began the process of shifting attention away from the traditional question 'how risky are the data for release?' towards the more critical question 'how might re-identification occur?' The data situation approach that we take here builds further on this and broadens our understanding to include the actions of other key agents, other data within the environment and previously-neglected considerations such as the importance of governance processes. The basic premise is that you cannot guard against the threat to anonymisation unless you have a clear idea of what it is you are guarding against and this requires considering both data and environment.

What this means for you is that your assessment and management of re-identification risk should include reference to all the components of the ADF, including your data, other external data sources, legitimate data use and potential misuse, governance practices, and your legal, ethical and ongoing responsibilities. The ADF is a total system approach, and consists of ten components:

1. Describe your data situation
2. Understand your legal responsibilities
3. Know your data
4. Understand the use case
5. Meet your ethical obligations
6. Identify the processes you will need to assess disclosure risk
7. Identify the disclosure control processes that are relevant to your data situation
8. Identify who your stakeholders are and plan how you will communicate
9. Plan what happens next once you have shared or released the data

[7] Some privacy models such as differential privacy and k-anonymity do attempt to assess and control risk by comparing it to some theoretically parameterised environment – there is however nothing intrinsic in these models that requires engagement with the actual data environment.

10. Plan what you will do if things go wrong

We will not say anything more here about these components as they are covered in some detail in chapter 3. What we will do is make explicit the five principles upon which the ADF is founded:

1. You cannot decide whether data are safe to share/release or not by looking at the data alone.
2. But you still need to look at the data.
3. Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data.
4. Zero risk is not a realistic possibility if you are to produce useful data.
5. The measures you put in place to manage risk should be proportional to the risk and its likely impact.

Let us consider these principles in a little more detail.

1. ***You cannot decide whether data are safe to share/release or not by looking at the data alone:*** This principle underpins the data situation approach outlined above, where risk is seen as arising from the interaction between data, people and (the soft and hard) structures that shape that interaction such as national policies on data sharing and access, the legal framework, IT systems, governance practices, cultural attitudes to data sharing and privacy etc.
2. ***But you still need to look at the data:*** You need to know your data – which means being able to identify the properties of your data and assess how they might affect risk. This will feed into decisions about how much data to share or release, with whom and how.
3. ***Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data:*** You may wonder why we talk about the need to balance data utility with data safety in the anonymisation process. It is easy after all to think about anonymisation only in terms of producing safe data but if you do that you may well be taking a risk for no benefit. Remember, anonymisation is a means inseparable from its purpose of sharing or releasing data. Let us consider this further:
    - On the issue of data utility – there is little point in releasing data that do not represent whatever they are meant to represent. There are two possible outcomes that arise from low utility and neither are happy ones: (i) the data are of little or no use to their potential users and you will have wasted your time and resources on them, or (ii) the data could lead to misleading conclusions which might have significant consequences if, for

example, the data are used to influence thinking or to make decisions which determine an outcome.
    - On the issue of data risk – low utility data may still retain some re-identification risk but in the absence of demonstrable utility you will lack any justification for taking that risk.
4. ***Zero risk is not a realistic possibility if you are to produce useful data:*** This is fundamental. Anonymisation is about risk management, nothing more and nothing less; accepting that there is a residual risk in all useful data inevitably puts you in the realms of balancing risk and utility. But this is the stuff of modern life – the trade-off of individual and societal level benefits against individual and societal level risks. This also brings into focus the issue of stakeholder engagement; there is no agreement on how to have a conversation with data subjects and the wider general public about this issue and there are not unfounded concerns about causing unnecessary worry by drawing attention to confidentiality risks. At the same time, it is worth recognising that people are capable of balancing risk and utility in much of their daily lives whenever they cross a road, drive a car etc.
5. ***The measures you put in place to manage risk should be proportional to that risk and its likely impact:*** Following principle 4, the existence of risk is not *a priori* a reason for withholding access to data. However, a mature understanding of that risk will enable you to make proportionate decisions about the data, who should have access and how. So for example:
    - If data are detailed and/or sensitive it would be proportionate for you to look to control the 'who and how' of access by, for example, limiting access to accredited users working in a secure lab facility.
    - If the data are of minimal detail and not sensitive then limiting access to a secure setting is likely to be disproportionate and it would be better to consider a less restricted access option.

1.3 Structure of this book

In this chapter we have introduced some of the core concepts relevant to our approach to confidentiality and anonymisation. We have also provided a top level overview of the Anonymisation Decision-making Framework, explaining both the thinking behind it and the principles on which it is founded. The ADF we have said is a generic approach to the process of anonymisation which will help you to identify and address the key factors relevant to your particular data share or release situation.

In the next chapter we define anonymisation in much more detail and bring together the ideas and concepts required to understand and adopt the ADF.

In chapter 3 we present the ADF, working through each component in detail. The approach taken is practical with worked examples and advice on how to operationalise each component of the framework. As we have prioritised accessibility over precision and completeness, some of the more technical aspects of disclosure risk assessment and control (for example synthetic data generation) are necessarily passed over but in many cases these are unnecessary and when they do prove useful it is generally better to work with an expert on their application. As with any complex topic there is always more to understand; this is an active research area and so the underlying science itself is still in development. You will see that we have made liberal use of footnotes. Our intention is that the book can be read and the framework understood without paying any attention to the footnotes at all – they are there for those who may want more detail.

Chapter 2 About Anonymisation

2.0 Introduction

In this chapter we will consider in depth what anonymisation is and the elements that make up best anonymisation practice. We will consider how anonymisation relates to other concepts, and how it is embedded in legal and technical ideas and practices.

2.1 Anonymisation and the law

Anonymisation is a process to allow data to be shared or disseminated ethically and legally, thereby realising their huge social, environmental and economic value,[8] whilst preserving confidentiality.[9]

An often misunderstood point is that anonymisation concerns keeping data *confidential;* it is not primarily about *privacy*. Privacy is a difficult to define, somewhat amorphous concept that implicates psychological notions like identity and autonomy, and depends on a locus of control which is contextualised by cultural norms. Confidentiality on the other hand relates directly to the collection and storage and transmission of information. Anonymisation processes may have privacy implications but they operate by maintaining confidentiality.

Fienberg (2005) defines confidentiality as 'a quality or condition accorded to information as an obligation not to transmit that information to an unauthorized party'. More specifically, in the context of personal data, the confidential matter is that these data relate to a particular person.[10] As a data subject, I do not (usually) mind if a data user can see that there exists a person that has my attributes but I am much more likely to object if they can ascertain that that person is me. The whole exercise of anonymisation is premised on that distinction.

---

[8] Often referred to as the triple bottom line (Elkington 1997).

[9] In some circumstances, non-anonymised data can also be shared but anonymisation usually makes sharing easier and wider dissemination of personal data without first anonymising it is usually not possible at all.

[10] Note this interpretation of confidentiality is specific to personal data. For example, in the sentence 'the details of company X's new wonder product were confidential' it is the details of the product that are confidential not the identity of company X.

All organisations collect some information from their clients/service users/members/employees as part and parcel of their activities and increasingly they share or even sell (at least some of) the data they collect. When this information relates directly to those persons as individuals then it is termed personal data.

In the UK, the law most relevant to personal data and their anonymisation is the 1998 Data Protection Act (the DPA), which enacted the 1995 European Data Protection directive. Other legislation that pertains to particular datasets and data sources such as the *Statistical Registration and Services Act* (2007) for official statistics, the *Commissioners for Revenue and Customs Act* (2005) for HMRC data and the *Census (Confidentiality) Act* (1991) for UK Census data, is also pertinent for anonymisation, but still approaches it via the notion of *identifiability* which is central to the DPA's concept of personal data.

In 2018, the new European Data Protection Regulation will come into force.[11] This will be a significant change in both the content of the legislation and how it is enforced. However, both the textual definition of what personal data is and the functional role of anonymisation in data protection appear to be unaltered by the new regulation. For the remainder of this book, we focus on the UK's Data Protection Act as the legislative framework in which anonymisation is deemed to take place. We will revisit this as part of our regular updates of this book.

The DPA defines personal data as:

> 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.[12]

The living individual[13] in question is often called the *data subject*.[14]

---

[11] At the time of going to press the UK had just voted to leave the EU. The impacts of this are obviously unclear at this stage. However, it seems likely that the UK will need to adopt compatible legislation

[12] We will discuss the important notion of the *data controller* in section 2.1.2 below.

[13] It often surprises people that the data for deceased individuals are not covered by data protection legislation. However, other pieces of relevant legislation such as the Human Rights Act and the Statistics and Registration Services Act do not make this distinction and in general it is prudent to not base anonymisation and data sharing policy on it.

[14] Often, particularly in the context of censuses and surveys, you will also see the word 'respondent' used. For individual-level data the two may be synonymous but sometimes a respondent will also provide information that is directly or indirectly about 3rd parties, such as members of the same household. So one can be a data subject and not a respondent and *vice versa*.

It is the second of these clauses - especially the crucial phrases 'other information' and 'is likely to' – that gives anonymisation its inherent practical complexity. But at the conceptual level, there are two simple questions that one needs to answer in order to understand whether data are personal or not: are the data about[15] people and are people identifiable within the data? The combinations of answers to those questions give us four possible types of data as shown in Table 2.1.

About People	**Non-Identifiable data**	**Identifiable data**
Yes	Anonymised Data	Primary Personal data
No	Apersonal Data[16]	Secondary Personal Data

Table 2.1. Four types of data depending on whether they are about people (or not) and whether they are identifiable (or not).[17]

One thing that may not be immediately clear is what the category of secondary personal data is; how can data not be about people but still be personal? An example of such data is the data about fires that, in the UK, are managed by the Department of Communities and Local Government on behalf of the fire and rescue services. Despite not being directly about people, the key point to realise about these data is that fires happen in *places* – often people's homes – and places are in turn associated with people. So whilst the data are not *about* people they may *relate to* them. If there is a close association between a place and a person – for example, the place is the person's address – then the person can be associated with particular elements of the data, which therefore may be personal.[18]

---

[15] We use 'about' here rather than the DPA technical term 'related to' deliberately. Data are *about* people if the data units are people. Data could *relate to* a person without being *about* them; for example a database of registered cars relates to the owners of those cars but the data topic is not people so the data is not *about* people.

[16] We introduce the term *apersonal* here. The reader might immediately wonder why we are not using the more familiar term *non-personal*. The point is to distinguish between data that are not to do with people from those that are to do with people but have been anonymised so that they are non-personal. So apersonal data are always non-personal but not vice versa.

[17] A slight confusion can sometimes arise because, in this context, the terms *identified* and *identifiable* can be applied to both people and data. So we might say that some data are identifiable because a person could easily be identified within them.

[18] Deciding whether information that is not directly about a person, but is about some object with which they are associated, is personal or not can seem a little tricky. So the fact there was a fire at my house may well be personal data in respect to me but the fact that my house is 263 feet above sea level is not. The ICO refers to this as the *focus* of the information (Bourne 2015). Fires might be personal because they involve the actions of people and have consequences for them. The information about

An interesting point to note is that in terms of volume the vast majority of data stored in computer systems are of the *apersonal* type, i.e. not relating to people. Astronomical data, meteorological data, food nutrition data, bus timetables, seismological data, stress readings for the Humber Bridge and lists of endangered species are all examples of apersonal data. Such data are clearly non-personal, and have little to do with humans. But what about the other forms of non-identifiable data?

2.1.1 So, are anonymised data non-personal?

Or alternatively, are the anonymised data that have been derived from personal data still personal? The straightforward answer to this would appear to be no, surely they are not because no-one can be identified in them – is that not the point of anonymisation? Unfortunately, the situation is more complex than this. Usually, following anonymisation, the original personal data still exist and this means that (except perhaps for the coarsest of aggregate data) the data controller will still be able to identify individuals within the anonymised data (using the original data as a reference[19]) and therefore it would seem that on a literal reading of the definition of personal data (cf. page 8) the data must still be personal. There are two ways of resolving this paradox:

1. To say that the anonymised data are personal and therefore the question about whether to share or release them depends on whether the DPA provides another get-out (e.g. whether the share or release constitutes fair processing[20]).
2. To say that the anonymised data are personal for the original data controller but non-personal for other users of the data.

We hold the second of these positions as it directly ties the concept of anonymisation to the notion of the context of personal data (in this case, other sources of data that

---

the elevation of a house above sea level is squarely and solidly about the house. The focus of the information is the house not the owner.

[19] In some data situations, the original data is destroyed – for example there is some legal or regulatory obligation to destroy the data or perhaps a guarantee to respondents that that will happen. Usually, the purpose of anonymisation in such situations is to allow some data to be retained whilst remaining compliant. The phrase *irreversible anonymisation* is sometimes used in these situations (meaning that the data controller can no longer re-identify the data).

[20] Fair processing is defined within the DPA as processing that meets one of a set of conditions (and also one of a separate set if the data are classified as sensitive). See UK: Information Commissioner's Office (2016) pages 98-103 for more information.

users have access to) and makes a clean separation between the complexities of data protection, such as the (essentially ethical) question of fairness, on the one hand, and the (essentially technical) question of identifiability on the other.

Another reason for not favouring the first definition is that it leads to some unintended consequences. For example: suppose I am a researcher and I collect some data via an anonymous web survey (quite common these days). If the survey is properly anonymous then I will not be able to identify anyone in the resultant data and therefore the data are not personal, and I am not a data controller and I am free to publish the data. But organisation X which has access to auxiliary data may be able to use those data to identify individuals within the data I have published. In this interpretation of the DPA, I am not liable because I cannot identify anyone in the web survey data and therefore am not a data controller. This seems counterintuitive at best.

In short, if you are considering whether data are anonymised and therefore non-personal you can only answer that question in the context of a given perspective. If the data controller has other information that enables them to re-identify a person within dataset X but the user of the anonymised dataset does not (and is not likely to), the dataset X is personal data for the data controller but not the user.

This interpretation is also shared by the UKs data protection regulator the ICO. However, in other jurisdictions the first resolution of the paradox is favoured. In those jurisdictions, data are deemed personal in and of themselves, irrespective of their context, if they are identifiable by a data controller. This is most clearly picked up in particular in the Article 29 Working Party of data protection regulators' opinion on anonymisation:

> *Thus, it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data.* EU: Article 29 Data Protection Working Party; Opinion 05/2014 on Anonymisation Techniques page 9.

A final point is that anonymising personal data is itself a form of data processing and therefore should not be seen as an escape clause from the DPA; as with all processing of personal data it must be compliant with the DPA. On the surface this may seem slightly esoteric. However, the key point is that anonymisation is not an end but a means and that one cannot separate the process of anonymisation from its purpose (which invariably is sharing or releasing data).

To understand this, consider the following, fairly conventional, scenario. Organisation X wants to release an anonymised version of dataset D. They go through a rigorous anonymisation process (and for argument's sake let us say that we know that this was done perfectly). They release the anonymised dataset but retain a copy of the original data.

At the point the decision is made to release the data those data are contained within the originating environment where they are personal. This is because the data controller has the means (the originating data) to re-identify them. This situation does not change post release;[21] they are still able to re-identify the anonymised data. Therefore the anonymised data remain personal for them and the DPA still applies.

Now, given that we are assuming perfect anonymisation, most of the principles of the DPA are clearly met.[22] For example, principle 7 concerning data security is intrinsically met directly by the anonymisation process. Principle 5 will be met as soon as the purpose that the original personal data were collected has been achieved (and the original data are destroyed rendering the anonymised data non-personal for everyone) and principles 3, 4 and 6 can only be meaningfully applied to the original data. This leaves us with principles 1, 2 and 8. Principle 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*

is potentially relevant to any open data release as open data is normally globally published via the Internet and therefore available in all countries regardless of their DP laws and practices.

Principle 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*

---

[21] This will not usually be true with organisation to organisation shares. Once the data has been shared then usually it will be sitting behind the receiving organisation's fire wall and the data controller for the originating data will not have access and therefore cannot have reasonable means to re-identify the anonymised copy. However, this situation does not alter the substance of the argument that follows.

[22] For the full text of the principles and their intended interpretation see: http://www.legislation.gov.uk/ukpga/1998/29/schedule/1

appears to have the potential to be tricky since it is often the case that anonymising data for release will not be included in the list of original purposes at the point of collection. However, the key point here is that although the (anonymised) data have been released the essence of what makes the data personal has not been disclosed. In other words, although anonymised data have been released, personal data have not been disclosed. This may seem a little paradoxical but a simple test illustrates the point: *does any legal person exist after the release for whom the data are personal and for whom they were not personal before the release*. If the answer is no then no disclosure of personal data has happened and that will be the case in the scenario we are describing here.

This leaves us with principle 1 – which concerns the fairness of processing – and this we cannot avoid because as stated above the anonymisation (which cannot be separated from its purposes) is a form of processing. This means in practice that the data controller should have a justification under the DPA's Schedule 2 (a legal basis)[23] for the anonymised share or release. The justification for an anonymised share or release would usually be either: (i) it is necessary for administering justice, or for exercising statutory, governmental, or other public functions or (ii) that it is in accordance with the legitimate interests of the data controller or (iii) for the exercise of any other functions of a public nature exercised in the public interest by any person. In the vast majority of cases where release or sharing of anonymised data is being considered one of those justifications will apply.

The role of anonymisation in this processing is to ensure that the data subjects' legitimate interest in the confidentiality of their data does not override the data controller's legitimate interest in sharing or releasing the data. However, what should be clear here is that the whole argument rests on the anonymisation process itself being thorough and rigorous. If it is not then the risk that personal data are disclosed will be non-negligible.

2.1.2 User, processor, controller – the roles of the anonymisation process

Understanding your legal status in respect of particular data is important as it will help you establish clearly what your responsibilities are and those of any other

[23] And also a schedule 3 justification if the data are sensitive.

stakeholders during the anonymisation process. It may also be that the design of the process will affect the roles that different agents play.

Let us begin by looking at the conditions under which you are considered a data controller and a data processor. The DPA defines a data controller as

> *… 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*.

It may seem an obvious point but it is worth making explicit that there are two conditions in this definition:

1. That a data controller determines the purposes and manner in which the data are processed.
2. That the data are personal data.

A data controller has overall responsibility for the **why** and **how** of data processing activities.[24] These activities include (but are not limited to):

- Making the decision to collect personal data in the first place and determining the legal basis for doing so.
- Determining which items of personal data to collect and the purpose(s) for which the data will be used.
- Determining whether to disclose data, and if so, to whom.
- Determining the need for anonymisation given the data situation.

Under the DPA, there can be more than one data controller for a given personal data product. This situation arises where multiple parties either 'jointly' or 'in common' determine the purpose for which, and the manner in which, the personal data are processed. The term 'jointly' refers to the situation where data controllers act together and equally in the determination of the processing of personal data. The term 'in common' refers to the situation where data controllers share a pool of personal data, each processing their share independently of the other controllers.

In contrast to a data controller, a data processor does no more than process personal data in the way(s) decided by the data controller. Their processing activities may include for example storing the personal data, providing security, transferring them across the organisation or to another and indeed anonymising them.

[24]For a more comprehensive list of data controllers' activities please see the ICO's (2014a) *Data Controllers and Data Processors*: http://tinyurl.com/ICO-CONT-PROC [accessed 30/05/2015].

Finally, for a person to be a user of some data but neither a processor nor a controller then such data are must be non-personal for that person and if the data are about people then they must be anonymised.

Thinking about peoples' roles with respect to some data can help structure anonymisation decision making. Let us take as an example the Administrative Data Research Network (ADRN) which provides researcher access to linked UK government administrative data through secure research centres. The original data owners of these data (the government departments) are data controllers and (since the data may be linked across from different departments) they will often be controllers in common. The individual Administrative Data Research Centres that make up the network and their associated trusted third parties are data processors. The mechanism by which the data are processed and accessed has been determined by the ADRN itself but the data owners decide whether to use that mechanism for a particular researcher/project. The researchers access the data under highly controlled conditions which make the risk of re-identification negligible, and therefore, because the data are *functionally anonymous* for them, they are users.

Identifying whether an agent is a data controller (solely, in common or jointly), a data processor or user is not always straightforward. But identifying the agents involved in a given data situation and their (desired) roles can help you decide what anonymisation processes are necessary and who should conduct them.

2.1.3 De-identification and anonymisation

There is a lot of confusion between the two terms *de-identification* and *anonymisation* mostly arising from the fact that the former is usually a necessary but rarely sufficient component of the later.[25] Here, we describe the two terms and outline some of the underlying issues that have led to the confusion.

De-identification – refers to a process of removing or masking *direct identifiers* in personal data such as a person's name, address, NHS or other unique number associated with them. De-identification includes what is called *pseudonymisation*.[26]

[25] Unfortunately, in some writing on the topic, this is exacerbated by treating the two terms as synonymous.

[26] Pseudonymisation is a technique where direct identifiers are replaced with a fictitious name or code that is unique to an individual but does not of itself directly identify them.

Anonymisation – refers to a process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves doing more than simply de-identifying the data, and often requires that data be further altered or masked in some way in order to prevent statistical linkage.[27]

We can highlight further the difference between anonymisation and de-identification (including pseudonymisation) by considering how re-identification might occur:

1. Directly from those data.
2. Indirectly from those data and other information which is in the possession, or is likely to come into the possession, of someone who has access to the data.[28]

The process of de-identification addresses *no more* than the first, i.e. the risk of identification arising directly from data. The process of anonymisation, on the other hand, should address both 1 and 2. Thus the purpose of anonymisation is to make re-identification difficult both directly and indirectly. In de-identification – because one is only removing direct identifiers – the process is unlikely to affect the risk of indirect re-identification from data in combination with other data.

It should be noted that in the description of both processes (i.e. de-identification and anonymisation) the purpose is to make re-identification more *difficult*. Both de-identification and anonymisation are *potentially* reversible; the data environment in which you share or release data is of critical importance in determining reversibility. In other words, the data environment can either support or constrain reversibility which means you need to think very carefully about the environment in which you share or release data. For example, it may be entirely appropriate to release de-identified data in a highly controlled environment such as a secure data lab but not at all appropriate to release them more openly, for example by publishing them on the Internet. The classic example of a failure of the data protection process to take into account the data environment is the release on the Internet of search queries by AOL in 2006 (see footnote 2). These were pseudonymised, yet people were clearly identifiable via common sense inference, such as if someone persistently searches for the name of a non-famous individual, it is likely to be that person himself. We will

[27] Statistical linkage refers to a process that classifies pairs of records across different datasets as matched (deemed to correspond to the same population unit) or not matched (deemed not to correspond to the same population unit).

[28] These options are obviously suggested by the DPA's definition of personal data, although the DPA refers only to the data controller we are talking here about anyone who has the data.

be coming back to the notion of the data environment later but for a full discussion see (Mackey and Elliot 2013; Elliot and Mackey 2014).

2.2 Types of anonymisation

The term 'anonymisation' gets used in a variety of different ways and inevitable communication difficulties arise as a consequence. Elliot et al (2015) have identified four different usages:

1. Formal Anonymisation
2. Guaranteed Anonymisation
3. Statistical Anonymisation
4. Functional Anonymisation

2.2.1 Formal anonymisation

For data to be formally anonymised simply requires that *direct identifiers* (sometimes called *formal identifiers*) have been removed from the dataset or masked in some way. Direct identifiers come in five forms:

1. ***Intentional Unique Identifiers:*** These are serial numbers that have been created with the explicit intention of identifying a person and for linking transactions. They are often used in multiple contexts and usually are associated with a person across his or her lifespan. Examples are UK National Insurance Numbers and US Social Security Numbers.

2. ***Digitised Unique Biometrics:*** These are codifications of unique, or statistically very likely unique, characteristics of individuals, to be used intentionally as identifiers. Their use can be intrusive, and – because they are hard to disavow – are often used in security contexts. Examples include fingerprints, iris scans, gait recognition systems, DNA and handwritten signatures.

3. ***Associational Unique Identifiers:*** These occur where some object which itself has a unique identifier is (strongly) associated with a person. Examples are a telephone number (particularly a mobile phone number), credit card number, static IP address or car registration number. They are invariably non-permanent but can exist for a while. General Unique Identifiers or GUIDs, which are used by Windows OS to identify software components and indeed users, and which in some cases can be semi-permanent, also fall into this category.

4. ***Transactional Unique Identifiers:*** These are numbers which have been generated as part of some transactional process. They are not necessarily permanent. Examples are sessional cookies and dynamic IP addresses.

5. ***Functional Unique Identifiers*** **(FUIs):** This category is a borderline one. Technically, they are a form of indirect identifier. However, what distinguishes them is that they map onto the first part of the definition of personal data ('can be recognised from these data'). The most straightforward example of an FUI is full name and address. FUIs will almost always be constructed out of more than one piece of information. They will also usually include the possibility of data twins (it might be that there are two people called 'John Henry Smith' living at address X), but these will be rare enough that we can treat FUIs as if they are unique.

2.2.2 Guaranteed anonymisation

For anonymisation to be guaranteed and irreversible there must in effect be zero risk of an individual being identified within a dataset given whatever assumptions one wishes to underpin the guarantee. This is the meaning of anonymisation that is usually employed within the security engineering literature (Ohm 2010; Dwork et al 2006) and in particular through the theory of differential privacy, which aims to provide a *privacy guarantee* using algorithms that make very specific (and invariably extreme) assumptions about what a data user might already know about the population represented in the data.

Ohm (2010) asserts that one can have anonymised data or useful data but not both and if one regards anonymisation as an irreversible process then he is correct. It may not be immediately obvious that this is true. So you might think for example that heavily aggregated data are 'irreversibly anonymised'. However, a theoretical intruder who has almost complete knowledge of the population from which the aggregated data were drawn but who lacks one piece of information about one particular individual could utilise what they already know to discover the piece of information that they are lacking (this is called a subtraction attack which we will discuss further in section 2.3.4). You might argue that this is a contrived situation and we would entirely agree. The point here is not to suggest this approach is sensible – it is not – but rather to illustrate how Ohm's assertion is a logically necessary consequence of the notion that risk can be removed from the process. However, we would argue that anonymisation should not be considered from this absolute standpoint.[29]

[29] The distinction we aim at here is analogous to the distinction in European civil contract law between the duty to achieve a specific result (*obligation de résultat*), and the duty to use one's best efforts (*obligation de moyens*). If one has an *obligation de résultat*, then there is a specific state that one is

Let us consider Ohm's position further using a simple analogy: *we can have secure houses or usable houses but not both*. If we assume that by secure we mean absolutely secure, then this is true. An absolutely secure house would lack doors and windows and therefore be unusable,.[30] But that does not mean that all actions to make one's house *more* secure are pointless, and nor does it mean that proportional efforts to secure my house are not a good idea. The deadbolt on my door may not help if a burglar comes armed with a battering ram or simply smashes my living room window but that does not mean that my lock is useless, merely that it does not (and cannot) provide absolute security.

And so it is with data. One has to balance data utility with re-identification risk, of which there will always be some. Fortunately, the DPA does not require anonymisation to remove risk entirely, but rather demands that those sharing or disseminating data mitigate the risk of re-identification until it is negligible (UK: Information Commissioner's Office 2012a).

The problem with guaranteed anonymisation is that in order to achieve it, one usually has to so restrict the data that it is often rendered useless (as Ohm points out). For example, as Sarathy and Muralidhar (2011) demonstrate, when differential privacy techniques are applied to an analysis server,[31] the net effect is that meaningful queries to the differentially private database are no longer possible. This finding is unsurprising when one considers that plausible re-identification attacks and meaningful data analysis both require data that differentiates population units, which differential privacy is designed to prevent. This tension is present with all data-focused anonymisation processes and sadly guaranteed anonymisation also guarantees data with little or no utility.

---

contracted to bring about. If one has an *obligation de moyens*, one is obliged only to use best practice and due care to achieve a goal. Ohm writes as if the data controller, while anonymising, is under an *obligation de résultat* to produce an irreversible state of anonymity of the data, while we argue that the data controller is under an *obligation de moyens* to understand the methods of anonymisation, and the properties and context of the data, in order to employ his or her best efforts to prevent re-identifications from the data.

[30] Artist Rachel Whiteread's concrete cast of the complete interior of a house makes this point quite nicely: https://en.wikipedia.org/wiki/House_%28sculpture%29 [accessed 30/5/2016]. Of course, as a work of art, this had no need for utility!

[31] An analysis server is a data environment where users do not see the actual data but instead submit syntax for analyses which are then run and then the researchers are sent the output (possibly after checking).

So in general, guaranteed anonymisation is not practical if one wants to share useful data – there will always be some risk associated with that activity. Risk naturally suggests a statistical treatment and this brings us to the third type of anonymisation.

2.2.3 Statistical anonymisation

The notion of statistical anonymisation is tied into a technical field called *statistical disclosure control* (SDC)[32] which we discuss in more detail below. The basic tenet of SDC is that it is impossible to reduce the probability of re-identification to zero, and so instead one needs to control or limit the risk of disclosure events.[33] This brings the notion of anonymisation into line with other areas of business risk management. One accepts that our actions and choices, responsibilities and constraints are embedded in a complex world which is impossible to predict in detail so one gathers the best information one can and optimises one's decisions to maximise the expected benefits and minimise the risks.

One could argue that both formal and guaranteed anonymisation are simply special cases of statistical anonymisation. Formal anonymisation is a mechanism for reducing the probability of re-identification below unity and guaranteed anonymisation is a mechanism for reducing it to zero. However, someone who releases or shares data that relates to individuals should have two goals: (i) to release/share useful data and (ii) for those data to be in a form which protects confidentiality (and thereby privacy). It should be clear from the foregoing that formal anonymisation will fail to achieve goal (ii), while guaranteed anonymisation will fail to achieve goal (i). Statistical anonymisation recognises that there is a lot of ground in between these two extremes.

---

[32] In the US, SDC is referred to as statistical disclosure limitation (SDL); the title 'Statistical Disclosure Control' is most commonly used in Europe. You will also see the term *statistical confidentiality* used but this tends to be used to mean 'the set of processes by which statistical data are kept confidential' and is therefore more general.

[33] It should be noted here that disclosure control researchers distinguish between *identification* and *attribution* processes in a disclosure. The former indicates that agent X has found person Y in some (supposedly anonymised) data, the latter indicates that agent X has learnt something new about person Y. These two processes often co-occur but need not. This is somewhat confusing because the two processes are conflated in data protection law; thus in the *Anonymisation Code of Practice* the UK Information Commissioner says 'Note that "identified" does not necessarily mean "named". It can be enough to be able to establish a reliable connection between a particular data and a known individual' (p.21).

At this point it is worth introducing a hybrid technique called *k-anonymisation*. In some ways this method is an attempt to take the best features of the guaranteed and statistical approaches and combine them in a single method (which also combines risk assessment and control). Essentially, k-anonymisation works by guaranteeing that for a given set of key variables (X) there exists no combination of values (Xj) for which there are fewer than k data units; k is defined by the entity carrying out the anonymisation. The general principle is that if a user knows fewer than k individuals with the attributes Xj then precise re-identification is prevented. We will have more to say on k-anonymisation (and its companion concepts) later.[34]

2.2.4 Functional anonymisation

Unfortunately, assessing disclosure risk even with the simplest of data is far from trivial. Indeed, a whole research community has built up around the topic with its own journals and conferences. Much of the work in this field has focused on the statistical properties of the data to be released/shared, primarily because this aspect of the disclosure risk problem is by far the most tractable. A great deal of headway has been made; sophisticated statistical models have been developed which have at least facilitated identification probability assessments anchored in the properties of the data.

However, as several authors (e.g. Paass 1988; Elliot and Dale 1999; Mackey 2009; Mackey and Elliot 2013) have pointed out, despite the advances in statistical disclosure control we are at best basing our measurement on only some of the determinants of the risk. There is a range of other issues:

1. The motivation of somebody wishing to attack anonymised data in order to re-identify somebody within it (this will affect *what* happens and *how*).
2. What the consequences of a disclosure are (which will affect the motivations of an individual to attempt a re-identification).
3. How a disclosure might happen without malicious intent (the issue of *spontaneous identification*).
4. How the governance processes, data security and other infrastructure for managing data access affect the risk.
5. The other data/knowledge that might be linked to the data in question (without which disclosure/identification is impossible if the data have undergone de-identification).

[34] We refer a reader interested in the technical discussion to Samarati and Sweeney (1998) and Samarati (2001), the thorough critique by Domingo-Ferrer and Torra (2008) and the recent review in the context of privacy models by Domingo-Ferrer et al (2016).

6. Differences between the data in question and the other data/knowledge (often referred to as *data divergence*).

Bringing these considerations into the framework of statistical anonymisation creates the fourth type: *functional anonymisation*. This addresses the contextual factors which Mackey and Elliot (2013) refer to collectively as the *data environment*. And it is these concepts that we will be explaining in the course of this book.

Although we have presented functional anonymisation as a separate type it does in fact overlap with other types and specifically it still requires the technical know-how that characterises statistical anonymisation. We will return to functional anonymisation and the data environment in section 2.4 after introducing some of the complexities of statistical disclosure control in more detail.

2.3 Anonymisation and statistical disclosure control

Statistical disclosure control is a complex topic and it is not our intention here to attempt to give a full airing to all the possibilities and nuances. If you want to dig deeper we would recommend you read one of the recent field summaries (Willenborg and De Waal 2001, Duncan et al 2011 or Hundepool et al 2012). Here, we sketch the ideas that are most important and useful for the anonymisation practitioner.

2.3.1 Building disclosure scenarios

A key component of a well-formed SDC exercise is the development of disclosure scenarios to ground risk analysis, specifying the risks semi-formally. Put simply, until you know what *could* happen, you are stuck with only a vague idea that the data are risky, and quite apart from being a stressful state of affairs this does not get you anywhere in practical terms.

Broadly speaking there are two types of disclosure risk: inadvertent disclosure and disclosure occurring through deliberate action.

Inadvertent disclosure and spontaneous recognition

A simple example will suffice to illustrate the notion of spontaneous recognition. Living next to me is a young married couple – very young in fact, both are sixteen. Unfortunately, the woman dies in childbirth leaving the man a 16-year-old widower with a baby.

Putting aside the sadness of this story, we do not suppose we will get many naysayers if we assert a belief that this combination of a small number of characteristics is extremely rare. Why is that? Well, we all have an intuitive knowledge of the population, biased perhaps by our own circumstances but reliable enough to enable us to assert with confidence that 16-year-old widows are unusual, 16 year old widowers are likely to be very rare and 16 year old widowers with a young child even more so. Might there be a good chance that my neighbour is the only one in the UK, or at least in my area?

Now suppose that I am using a de-identified dataset and I come across a record of a sixteen year old widower with a young child who lives in my area. I might assume that it is my neighbour. This then is spontaneous recognition: the unmotivated identification of an individual in a dataset from personal knowledge of a small number of characteristics.

Of course such judgements are subjective and subject to availability bias, overconfidence effects and other forms of cognitive bias. So claims to have found someone can easily be misjudgements. Let us look at the example a little more objectively. At the 2011 census there were seven 16-year-old widowers in the UK. So my neighbour is not unique but an example of a rare combination of attributes. However, if one added in the fact that this person has a young child and included any sort of geographical indicator then the probability of the data actually singling out my neighbour would be quite high. So, theoretically, the risk of inadvertent and accurate recognition is non-zero.

However, bear in mind here that the presumption of this scenario is that the recognition is inadvertent, and the lack of any prior motivation substantially reduces the privacy risk for two reasons.

Firstly, the example is not so much of me finding a needle in a haystack but just happening to sit on one. The dataset has to be configured in such a way that the unusual combination of characteristics that my neighbour has appears simultaneously in my software window as I am browsing the data. For a large dataset the likelihood of such an event will be pretty low.

Secondly, having recognised my neighbour, what am I going to do? If I decide to act on my discovery then this is no longer simply a case of spontaneous recognition but a particular type of deliberate attack called fishing. If on the other hand I do nothing then this might be a ‘so what?’ situation, in which no harm befalls my neighbour,

with minimal privacy impact. The meaning of the recognition will partly depend on what the dataset is about; if it is a dataset of criminals or sufferers from sexually-transmitted diseases then simply being in the data is sensitive and me finding my neighbour in there might matter a lot. On the other hand, if it is a random sample of some country's population then maybe spontaneous recognition matters less.

Other factors which will indicate whether one need be concerned with spontaneous recognition are the size of the dataset, whether the user has response knowledge and who the users are.

Dataset size can have a counterintuitive effect. A smaller dataset effectively decreases the size of the haystack so it increases the likelihood of coming across someone (if they are in there).

Response knowledge; we will talk about this in more detail shortly. But simply put if I know you are in the dataset then I am more likely to spot your combination of characteristics and more likely to assume that it is you if I do so.

Who the users are; with open data the users are potentially the whole world and if it is high utility data then the actual user base might be very large. The larger the user base the more likely a spontaneous recognition event will be. In some data situations there might be a relationship between the user and the data subjects (for example an academic doing research on student data) and this can increase the risk.

One data situation where all three of these factors can come into play is the in-house survey and in particular the staff satisfaction surveys that are now commonplace in all sectors. The datasets tend to be small and drawn from a particular population with which the users of the data (the organisation's management) have a relationship. The users know that many (or even all) members of staff will be in the survey. In this type of data situation spontaneous recognition can be a serious possibility.

Deliberate attacks and the data intruder

In SDC, the agent who attacks the data is usually referred to as the *data intruder*.[35] As soon as you consider such a character as a realistic possibility rather than a shady abstraction, several questions immediately arise such as who might they be and what might they be trying to achieve by their intrusion? Considering such questions

[35] Other terms that are used are 'the attacker', 'the data snooper' and 'the adversary'. These are synonymous.

is an important first stage in the risk management process. Elliot and Dale (1999) have produced a system of scenario analysis that allows you to consider the questions of who, how and why. This method involves a system of classification which facilitates the conceptual analysis of attacks and enables you to generate a set of *key variables* that are likely to be available to the data intruder. We have further developed this system for the purposes of the Anonymisation Decision-Making Framework. The classification scheme is as follows:

INPUTS

- ***Motivation:*** What are the intruders trying to achieve?
- ***Means:*** What resources (including other data) and skills do they have?
- ***Opportunity:*** How do they access the data?
- ***Target Variables:*** For a disclosure to be meaningful something has to be learned; this is related to the notion of sensitivity.
- ***Goals achievable by other means?*** Is there a better way for the intruders to get what they want than attacking your dataset?
- ***Effect of Data Divergence:*** All data contain errors/mismatches against reality. How will that affect the attack?

INTERMEDIATE OUTPUTS (to be used in the risk analysis)

- ***Attack Type:*** What is the technical aspect of statistical/computational method used to attack the data?
- ***Key Variables:*** What information from other data resources is going to be brought to bear in the attack?

FINAL OUTPUTS (the results of the risk analysis)

- ***Likelihood of Attempt:*** Given the inputs, how likely is such an attack?
- ***Likelihood of Success:*** If there is such an attack, how likely is it to succeed?
- ***Consequences of Attempt:*** What happens next if they are successful (or not)?
- ***Effect of Variations in the Data Situation:***[36] By changing the data situation can you affect the above?

This approach in scoping the who, why and how of an attack owes as much to criminology as it does to technical risk analysis.

[36] Recall that a data situation concerns the relationship between some data and their environment. We discuss this in more detail below.

In order to make sense of this scenario-classification scheme you need to understand a set of basic concepts: key variables, data divergence, and response knowledge. We will go through each of these in turn explaining how they fit into the scenario classification scheme as we go.

Key variables

The pivotal element in the scenario analysis is the identification of the key variables.

These are essential for the intruder to achieve re-identification and allow association of an identity with some target information. Key variables are those for which auxiliary information on the data subjects is available to the data intruder and which provide a 'hook' into the target dataset, allowing individuals to be matched. See Figure 2.1 for a schematic view of how this works. Ideally, from the intruder's point of view, the coding method of a key variable must be the same on both the attack and target datasets, or the two must at least be harmonisable.

Essentially, there are four sources of auxiliary information: (i) datasets containing the same information for the same (or sufficiently similar) population, (ii) information that is publicly available (e.g. in public registers or on social media), (iii) information obtained from local knowledge (e.g. house details obtained via an estate agent or by physical observation), and (iv) information obtained through personal knowledge (e.g. things I know about my neighbours or work colleagues).

There is obviously a terminological overlap between the notion of a key variable and that of an indirect identifier. The distinction is that a key variable is specific to a particular scenario (for example a particular combination of datasets) whereas the term indirect identifier is focused on the dataset itself and which variables could be used as identifiers in any scenario. So in effect the set of indirect identifiers is the set of all possible key variables across all possible scenarios. But – and this is critical – one would very rarely (if ever) encounter a situation where one considered all potential indirect identifiers simultaneously as most scenarios will only involve a subset – the key variables for that scenario.[37]

[37] We note that the k-anonymity literature uses the term *quasi-identifiers* to refer to both key variables and indirect identifiers which in our experience does sometimes lead to some confused thinking by practitioners; so the terminological separation is not just a matter of semantics.

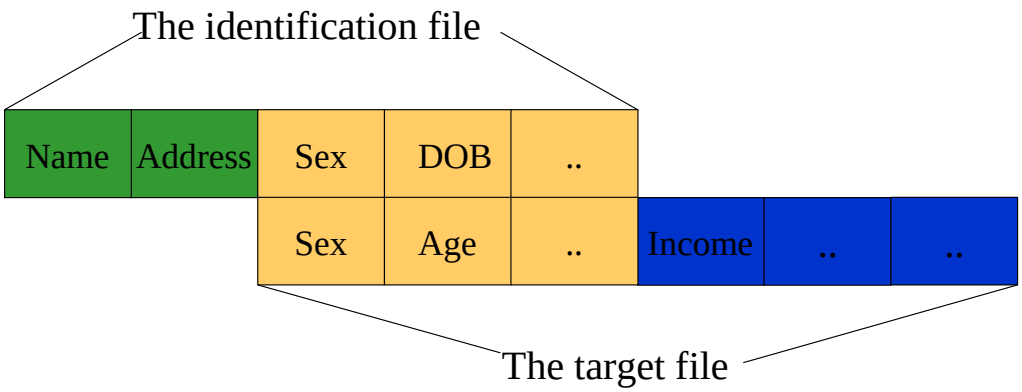


Figure 2.1: An illustration of the key variable matching process leading to disclosure. From Duncan et al (2011).

Data divergence

Another critical point in the scenario framework is consideration of data divergence. All datasets contain errors and inaccuracies. Respondents do not always supply correct data. Interviewers make mistakes in recording. Data coders transcribe incorrectly. Data items are missing. Missing or inconsistent values may be imputed using methods with no guarantee of accuracy. Data may be months or possibly years old before they are disseminated and characteristics will have changed since the data were generated. This is true of the target dataset as well as the auxiliary information held by an intruder. The combination of these will introduce the possibility of error into any linkage.

Collectively, we refer to these sources of 'noise' in the data as *data divergence*. The term refers to two situation types (i) *data-data divergence*, or differences between datasets, and (ii) *data-world divergence*, differences between datasets and the world. In general, both types can be assumed to reduce the success rate of matching attempts. However, where two datasets diverge from the world in the same way, which we call *parallel divergence*, then the probability of correct matching is unaffected. This would be the case, for example, if a respondent has lied consistently or when two datasets both have out-of-date but identical data.

Taking data divergence into account in a coherent way is complicated and it tends to mean that orthodox risk measures overestimate the risk (given the scenario). Elliot and Dale (1998) estimated that the effect in their particular study was to reduce the number of correct unique matches by as much as two thirds. This is one reason why it can be important to carry out intruder tests as well as data analytical risk assessments.

Notwithstanding the above remarks, a paradox of data divergence is that it is not a reliable protection of confidentiality. Firstly, for any particular individual-record pair there may be no divergence at all. Secondly, analysts are getting increasingly sophisticated in dealing with linkage in the face of 'fuzziness' – when we talk through the process of doing a penetration test in chapter 3 you will see that we attempt to tackle that issue.

So the best way to think about data divergence is that it provides you with a little extra protection – a margin of error rather like the reserve in a car's petrol tank – it is good as back up but not to be relied on.

Response knowledge

At its simplest level the issue of response knowledge can be captured by a single question: 'Do you know I am in the data?' If the answer to that is 'yes' then you are said to have response knowledge of me in respect of those data.[38] In that case, one key element of uncertainty, whether the person is even in the data at all, is removed. In practice, response knowledge can occur in one of two ways:

1. The intruder knows that (a) the data correspond to a population and (b) the target is a member of that population.
2. The intruder has ad hoc knowledge about a particular individual's presence in the data (e.g. my neighbour told me that she had been surveyed).[39]

The second is relatively simple to understand and is particularly pertinent to an open data situation. The first is more complex as 1(b) can be nuanced. Consider a hypothetical anonymised dataset of the members of the Bognor Regis Bicycle Club. Straightforwardly, I could know that my target is in the club and therefore in the dataset. That is clear cut response knowledge but I could have other information about you which falls short of full response knowledge but is nevertheless

[38] Of course you might be wrong – perhaps the information which tells you I am in the data is out of data or misattributed. Technically, response knowledge should be called something like 'beliefs about particular population units' presence in a particular dataset' but it is not a very reader friendly formulation. This is part of a more general issue of data divergence and applies even to direct identifiers (I might think I know your name and address but I could be mistaken). We will discuss this general issue in more detail shortly.

[39] Another theoretical possibility is that the intruder has inside knowledge of the data collection process. This would imply a complex security breach involving a situation where the intruder did not gain access to the raw data but did have access to an anonymised version of the data. Although this should not be discounted it is obviously quite obscure and the key problem here is the security breach, not the anonymisation problem.

informative. I could know that you live or work in Bognor Regis or that you are an avid cyclist or perhaps that you are a compulsive club-joiner. All of these constrain the *super-population* that contains the Bognor Regis Bicycle Club population and that in turn increases the effective sample fraction.[40] As we will see in chapter 3 the sample fraction is an important element of the risk.

2.3.2 Uniqueness

The example above brings us to *uniqueness,* one of the fundamental concepts in disclosure risk assessment, which underpins much of the research on disclosure risk analysis. A record is unique on a set of key variables if no other record shares its combination of values for those variables.

For disclosure risk purposes we need to examine two types of uniqueness on a set of key variables: population uniqueness—a unit is unique in the population (or within a population data file such as a census); and sample uniqueness—a sample unit is unique within the sample file.

A simple example – using just two variables – should clarify the relationship. Imagine that there are twenty people living in a village and we have some data on their ages and sexes as shown in Table 2.2. Now if you peruse the table you will see that there are two people who have a unique combination of characteristics in the data, Jeffrey Magnolia and Jessica Black. Now imagine that we take a 50% random sample of this population. One possible sample is shown in Table 2.3; we have also de-identified this sample by replacing the name with a sample ID.

If you peruse this table you will see that we have 4 records that are unique in the sample; the ones with sample IDs 3, 5, 9 and 10. Only one of these (number 10, the one corresponding to Jeffrey Magnolia) is actually unique in the population. The other sample uniques have *statistical twins* in the population – units sharing the same attributes. So, for example, we cannot tell whether record 9 corresponds to Jane Azure or Julia Beige. Jessica Black who is unique in the population is not in the sample.

[40] The sampling fraction is the proportion of a population to be included in a sample. It is equal to the sample size divided by the population size.

Name	Age group	Sex
Johnny Blue	0-16	Male
Jenny Blue	0-16	Female
Sarah White	0-16	Female
Sam Brown	0-16	Male
Julia Black	0-16	Female
James Green	17-35	Male
Peter Grey	17-35	Male
Jemima Indigo	17-35	Female
Jim Blue	17-35	Male
Joshua White	17-35	Male
Joan White	17-35	Female
Jill Brown	17-35	Female
James Brown	36-64	Male
Jessica Black	36-64	Female
Joe Orange	36-64	Male
John Black	36-64	Male
Jacqui Purple	65+	Female
Julie Beige	65+	Female
Jane Azure	65+	Female
Jeffrey Magnolia	65+	Male

Table 2.2 Ages and sexes of all people living in 'Anyvillage'

Sample ID	Age group	Sex
1	0-16	Male
2	0-16	Male
3	0-16	Female
4	17-35	Male
5	17-35	Female
6	17-35	Male
7	36-64	Male
8	36-64	Male
9	65+	Female
10	65+	Male

Table 2.3 Ages and sexes of a 50% sample of the people living in Anyvillage

In one form or another, these two concepts – sample and population uniqueness – form the basis of many of the disclosure risk assessment methods for microdata (files of records about individuals). If a unit is population unique then disclosure will occur if an intruder *knows* it is population unique. Much of the methodology in this area concerns whether sample information can be used to make inferences about population uniqueness.

The simplest inference is that given the sample file, if a record is not unique in the sample file it cannot be unique in the population, while a record that is unique in the population will be unique in the sample if it appears at all. This will not get the intruder very far but as we will see later not all sample uniques are the same.

2.3.3 Attribution and identification

Technically, statistical disclosure can occur through one of two distinct processes: *re-identification* and *attribution*. Re-identification (or identity disclosure) is the process of attaching an identity to some data. Attribution (or attribute disclosure) is the process whereby some piece of information is associated with a population unit.

The two processes can sound very similar but the distinction is quite important in terms of how disclosure risk is assessed for different types of data. In essence, identification means we find a person; attribution means we learn something new about them. Although the two processes often occur simultaneously, they can in fact occur separately.

Formally, a disclosure happens when an attribution is made, not when a re-identification happens. Accurate re-identification typically (but not always) leads to attributions, but attributions can happen without re-identification. For example, if I know that one of five of the records in a dataset corresponds to you and all of those records are of bank managers then I now know you are a bank manager even though I have not associated your identity with a particular record.

In the UK, the ICO has made it clear that reliable attribution does count as re-identification in their interpretation of the DPA:

> *Note that 'identified' does not necessarily mean 'named'. It can be enough to be able to establish a reliable connection between particular data and a known individual.* UK: Information Commissioner's Office (2012a p 21).

This might seem a little confusing, but, in the above example, if I have learnt that you are a bank manager then in effect the datum 'is a bank manager' has been associated with you, and confidentiality has been breached. I may not know which of the five records are yours and therefore I cannot name the record that belongs to you. However, I do not necessarily need to be able to do that in order to find out something about you from the data.

We will look at the mechanism for this in a little more detail in a moment. For now the key takeaway message is that any form of statistical disclosure counts as re-

identification from the point of view of the DPA. So making your data non-disclosive (in the context of its environment) will ensure that your processing is compliant with the DPA.

2.3.4 Types of attack

Re-identification attacks through linkage

Re-identification through linkage is the canonical form of disclosure risk. The presupposition is that a data intruder has access to some information which contains formal identifiers for population units and a set of *key variables* which are also present on the target dataset. The key variables are then used to link the identifiers to the target information —in principle, this could be any information not already known to the data intruder but in practice, in the scenario framework, we assume that the information has some value in terms of their goal. This is shown schematically in Figure 2.1.

Formal risk assessment for microdata[41] releases usually requires us to understand the probability of the data intruder being able to make such linkages correctly.

Attribution attacks

Consider the tables of counts shown in Table 2.4. Suppose the population represented in this table is everyone at a workshop I am attending. Over drinks, I overhear someone saying that they earned over two million pounds in the last quarter. Now I can infer that person is a lawyer. This is positive attribution — the association of the attribute 'is a lawyer' with a particular person. Conversely, if I hear somebody talking about their students, I can infer that they do not have a high income. This is a negative attribution — the disassociation of a particular value for a variable from a particular population unit.

[41] Datasets of records of individual population units.

Occupation	Annual income			Total
	High	Medium	Low	
	>250K pa	40-250K pa	<40K pa	
Academics	0	100	50	150
Lawyers	100	50	5	155
Total	100	150	55	305

Table 2.4: Table of counts of income levels for two professions from hypothetical population. From Duncan et al (2011).

Note that, in effect, association and disassociation are different forms of the same process, attribution arising from zeroes in the dataset. The point to note is that the presence of a (non-structural)[42] zero in the internal cells of a table is potentially disclosive.

Subtraction attacks

Now consider Table 2.5. The population in this table differs from that in Table 2.4 in one respect—we have one highly paid academic. Given this table, I can no longer make the inferences that I could from Table 2.4 (at least not with certainty). However, what about myself? I am a member of the population represented in the table and we can assume that I know my own occupation and income! Suppose that I am a highly-paid academic. Given this extra piece of knowledge, I can subtract 1 from the high income academic cell in the table, which then reverts to Table 2.4 and I am back to the situation where I can make disclosive inferences from overheard partial information about particular individuals.

Occupation	Annual income			Total
	High	Medium	Low	
	>250K pa	40-250K pa	<40K pa	
Academics	1	100	50	151
Lawyers	100	50	5	155
Total	101	150	55	306

Table 2.5 Table of counts of income levels for two professions from hypothetical population. From Duncan et al (2011).

[42] A structural zero occurs when a combination of attributes is impossible. For example, the number of three year old married people would, in the UK, produce a structural zero because of UK law. Non-structural zeroes appear where there are possible combinations of attributes which happen not to be instantiated. So there might happen to be no sixteen year old married people in Anyvillage in my data but the existence of such a person is possible.

We can extrapolate further. Consider a situation where I have complete information (in terms of the two variables contained in Table 2.5) about multiple individuals within the population. In effect such information represents a table of counts of the subpopulation of the individuals for whom I have complete knowledge. On the assumption that identification information is available for both that subpopulation and for any additional information I gain through overheard conversations (or other sources of data), I can subtract the whole of that table from the population table before proceeding. In principle, this could lead to more zeroes appearing in the residual table. The 'low-paid lawyers' cell would be particularly vulnerable to further subtraction and this illustrates a further crucial point: whilst zero counts are inherently disclosive, low counts also represent heightened disclosure risk, because they make it easier to obtain sufficient information external to the aggregate table to enable subtraction to zero than high cell counts.

Inference attacks

Beyond the subtraction to zeroes there is another sense in which low cell counts constitute a risk. Consider again Table 2.5. Now recall that it is not possible, without external information about the population represented in the table, to make inferences about any given individual with certainty. However, imagine again that I overhear someone at the workshop boasting about their high income. Whilst I cannot say with certainty that this individual is a lawyer, I can say so with a high degree of confidence. From the table, the conditional probability that a randomly selected person is a lawyer given that they are a higher earner is greater than 0.99.

This is inference – the capability of a user of some data to infer at high degrees of confidence (short of complete certainty) that a particular piece of information is associated with a particular population unit. Such inferential capacity could also in principle be derived from statistical models and other statistical output.

Depending on circumstances, this inferential knowledge may be good enough to meet the data intruder's goals. Deciding in any categorical sense what level of certainty of inference constitutes a problem is impossible. The best approach for dealing with this issue is to understand whether an inference at a particular level would be a success for the intruder and then whether that inference would cause harm to a data subject. This reiterates the necessity of well-formed disclosure scenarios.

Differencing attacks

A difference attack is possible with variables for which there are multiple different plausible coding schemes for a variable, where the categories in those coding schemes are not nested but instead overlap. This situation may occur where there are separate requests for tables or maps with different codings potentially allowing more information to be revealed about those in the overlaps than intended from a single table. Although it could happen with any variable the issue most commonly comes up with Geography.

The end result of this is that whilst a table may be considered safe in isolation, this may not be the case for multiple tables when overlain with one another.

Complex attacks

The attacks mentioned above are the simple ones. There are more complex operations that a sophisticated intruder can try, often with lurid names that can confuse and befuddle: *table linkage*, *mashing attacks*, *fishing attacks*,[43] *reverse fishing attacks* and so forth. It is outside the scope of this book to go into the details of these but suffice it to say that all of these involve bringing together multiple data sources. In practice if one covers the simple attacks then the complex ones also become more difficult to execute. However, you must also bear in mind that if you release multiple data products from the same personal data source into the same environment then you will be increasing the risk and you therefore need to proceed with caution. One way in which this comes up is where both microdata samples and aggregate whole population counts are released from the same underlying dataset. This is a common practice with censuses. To give a simple illustration, let us return to our hypothetical sample dataset in table 2.3 and add another variable, 'has cancer', to it (see Table 2.6 below). Now if I know a person who is Male and 65+ who lives in Anyvillage then I might suspect that it is case 10, but it is a 50% sample so I cannot be sure that my acquaintance is even in the data.

However, suppose that the data controller also publishes the Table 2.7 on its web site. On its own, the table looks fairly innocuous – but by combining this with the microdata to which the data controller has allowed me access I am able to ascertain

[43] Fishing attacks should not be confused with Phishing. Phishing is fraudulently obtaining personal authentication information (usually passwords) by pretending to be a third party (often a bank). A fishing attack on the other hand is the identification of an unusual record in a dataset and then attempting to find the corresponding entity in the world.

that my acquaintance has cancer. This example is obviously quite simplistic. With real data situations the interactions between different data products drawn from the same data source can be more subtle. To reiterate the take home message here: be very careful if you are considering releasing multiple data products from the same data source.

Sample ID	Age group	Sex	Has cancer
1	0-16	Male	No
2	0-16	Male	No
3	0-16	Female	No
4	17-35	Male	No
5	17-35	Female	No
6	17-35	Male	Yes
7	36-64	Male	No
8	36-64	Male	No
9	65+	Female	No
10	65+	Male	Yes

Table 2.6: Hypothetical 50% microdata sample of the people living in Anyvillage

	Age group				
Sex	0-16	17-35	36-64	65+	Total
Female	3	3	1	3	10
Male	2	4	3	1	10
Total	5	7	4	4	20

Table 2.7: Crosstabulation of people living in Anyvillage by age group and sex.

2.3.5 Types of formal disclosure risk assessment

Broadly speaking there are two types of disclosure risk assessment: Data Analytical Risk Assessment and penetration testing. The two approaches have complementary advantages and disadvantages.

Data Analytical Risk Assessment (DARA)

This is sometimes referred to as statistical disclosure risk assessment. It covers a large range of techniques from the very simple (counting uniques or identifying small cells) to more complex ones involving constructing statistical or computational models.[44] What they have in common is that they take the dataset in question as an

[44] We will not go into the details of the modelling approaches here and would refer the interested reader to Hundepool et al (2012) for a recent technical review.

analytical object, treating disclosiveness as a property of the data and attempting to identify the level of that property latent in the data.

Done well, DARA should be grounded in scenario analysis. However, even with this in place, there are several disconnects between the analysis and what would happen in a real attack; most importantly no external data are involved in DARA. Having said that, if the analyst is mindful that (no matter how sophisticated the techniques) they are only producing proxy measures for the real risk then DARA can be very informative.

In chapter 2 we will run through one approach that you can take to DARA.

Penetration tests

Another way of assessing disclosure risk, detailed in two of our case studies, is what we refer to as *penetration testing* (also known as *intruder testing*). The idea of penetration testing is to replicate what a plausible motivated intruder might do (and the resources they might have) to execute a re-identification and/or disclosure attack on your data.

The ICO have characterised what they refer to as a 'motivated intruder' as someone who is relatively competent, who has access to external data resources such as the internet and public documents, and is willing actively to make enquires to uncover information. They are not assumed to have specialist knowledge or advanced computer skills, or to resort to criminality. You can of course use a different set of assumptions about the type of knowledge skills and resources that an intruder has if to do so makes sense within your own scenarios.

There are essentially four stages to a penetration test: (i) data gathering; (ii) data preparation and harmonisation; (iii) the attack itself; and (iv) verification. The first stage tends to be the most resource-intensive whereas (ii) and (iii) require the most expertise. We go into these in more detail in chapter 2.

There are three core advantages of intruder testing as a risk assessment method compared to DARA approaches:

1. It mimics more precisely what a motivated intruder could do.
2. It will explicitly take account of data divergence.
3. It is based on real data gathering and real external data.

In other words it is *grounded*. Against this, it has one important disadvantage: it will be tied very tightly to one particular exercise and therefore does not necessarily

represent all of the things that could happen. This disadvantage is the flip-side of its advantages and indeed is an issue with all testing regimes: one trades off groundedness against generality and so in practice one should combine data analytical techniques with intruder testing rather than relying solely on either one.

2.4 Functional anonymisation and the data situation

The foregoing discussion should suffice to illustrate that anonymisation is a complex topic with many different components and that simply considering one aspect in isolation could lead to difficulties, and a non-functional solution.

Functional anonymisation considers the whole of the data situation, i.e. both the data and their data environment. When we protect confidentiality we are in essence hoping to ensure that anonymised data remains anonymous once it is shared or released within or into a new data environment and therefore functional anonymisation has to consider all relevant aspects of this situation.

We need to address the disclosure problem in this way because it is meaningless to attempt to assess whether data are anonymised without knowing what other information is or could be co-present. As we have seen, this is explicit in the definition of personal data in law and yet practitioners often attempt to judge whether data are personal or not using absolute criteria (the non-relative properties of the data themselves). This is based in part on the misapprehension that anonymisation can be absolute without mangling the data so badly that it has no utility whatever and in part on being overly focused on the data themselves. If anonymisation is to be a useful tool for data and risk management, one has to specify its circumstances. Thus the only sensible response to the question 'are these personal data?' is another question: 'in what context?' or more specifically 'in what data environment?'

How, then, might we formalise the notion of a data environment to allow such questions to be answered? Formally, we posit that a data environment is made up of four components: data, agency, governance processes and infrastructure.

1. ***Data***: What (other) data exist in the data environment?[45] How do they overlap with or connect to the data in question? This is what we need to know in

[45] Amongst all the challenges that anonymisation brings this question is probably the one that causes those who are responsible for it the most stress and lost sleep. At best, any answer to the question will be partial. The inexorable increase of the quantity of data 'out there' means this is necessarily so. However, it is important to keep this in perspective. Firstly, for nearly forty years data controllers

order to identify what data (key variables) are risky, and can be used for statistically matching one dataset with another thereby improving the conditions for statistical disclosure.

2. ***Agency***: We consider agents as capable of acting on and in the data environment. It may seem like an obvious point but it is one worth emphasising – there is no risk of a data confidentiality breach without human action or misdeed.[46]
3. ***Governance processes:*** We use the term here broadly to mean how users' relationships with the data are managed. This includes formal governance (e.g. laws, data access controls, licensing arrangements and policies which prescribe and proscribe user behaviour) through *de facto* norms and practices to users' pre-dispositions (e.g. risk aversion, prior tendency towards disclosure, etc.).
4. ***Infrastructure*****:** We use this term to consider how infrastructure and wider social and economic structures shape the data environment. Infrastructure can be best thought of as the set of interconnecting structures (physical, technical) and processes (organisational, managerial, contractual, legal) that frame and shape the data environment. Infrastructure includes information systems, storage systems, data security systems, authentication systems and platforms for data exchange.

A straightforward example of a data environment that can be described using all of these features is a secure data centre. It has data, data providers, a user community and context-specific physical, technical, organisational, and managerial structures that determine what data goes in, how data is stored, processed, risk-assessed and managed, the format in which data comes out, who the user community is, and how it can interact with those data. Data environments can of course be looser in form than a secure controlled data centre. An environment might be defined by regulation

---

have been releasing data that has been through anonymisation processes, resulting in only a small number of problems, almost all caused by very poor anonymisation decision making. Secondly, there are some simple things that you can do which will mean that you move beyond simple guesswork and that will put you firmly in the best practice camp. We will discuss these further in chapter 3, component 6.

[46] The development of AI and machine learning may soon make this categorical statement less certain. However, the question about when and whether non-human entities may count as agents in the sense that we employ here (and indeed more generally be included in humanity's moral universe) is clearly outside the scope of this book, and takes us to the heights (or depths) of philosophy. For the present, if we only concern ourselves with human agency, we will not be missing any pressing practical issues.

and licensing that allows (specific) users access to data under a licence agreement which stipulates what can and cannot be done with them. Such an environment cannot be as tightly controlled as the secure data centre environment, but it does allow for some control which is not present when, for example, data are published on the internet.

Environments exist inside other environments. The secure setting might sit within a bigger organisational data environment and the organisation in turn exists within the global environment. One of the aims of governance and security infrastructure is to prevent data leaking into or from these larger environments. A high level of confidence in security implies that correspondingly less attention needs to be given to the wider environment when considering risk. Obviously if you are publishing data openly then you will not have that luxury.

Now that you have an understanding of what a data environment consists of and might look like, you can begin to think about the notion of environment in relation to your own data products and how you might want to share and or release them. This brings us the concept of a data situation, a term intended to capture the idea of the relationship between data and their environment. What we are really interested in is in helping you to describe and understand your own data situation(s).

Data situations can be *static* or *dynamic.* In static data situations, the data environment is fixed, whereas in dynamic data situations it is subject to change. Any process of sharing or releasing data creates a dynamic data situation, as does a *de facto* change in the data's current environment (for example the relaxation or tightening of security processes). Once the share or release is complete then the environment fixes again and the data situation may revert to being static.[47]

By mapping the data flow from the point at which data are collected to the point after which they are shared or released you will be able to define the parameters of your data situation.

[47] Data situations can be static and dynamic, but we should also consider that data themselves can be static and dynamic too. Dynamic data are data that are being constantly updated through a data stream and a data stream is one type of dynamic data situation. The key point here is that in a dynamic data situation the data are moving relative to their environment. Dynamic data create a *de facto* dynamic data situation but so does the movement of static data.

So to reiterate, functional anonymisation is a process which controls disclosure risk by considering the totality of a data situation. We discuss this as a practical concept in chapter 3. For a deeper discussion of the theory see Elliot et al (2015).

2.5 Anonymisation solutions

In this section we review the various options you have to reduce the risk of disclosure from your data down to a negligible level; in other words to carry out functional anonymisation. These options fall into two groups, those focused on the data and those focused on the data environment. Normally you will need both. Before we move on to discuss the solutions in detail, we first want to discuss the unavoidable trade-offs that you will need to make as part of your anonymisation process.

2.5.1 Risk-utility and other trade-offs

Because anonymisation is about producing safe, usable data, we need to understand the trade-off between the two. Often the information that makes data risky is what makes it of interest to bona fide analysts. However, that is not always the case and as we will see in chapter 3, one of the important parts of functional anonymisation is considering the use case. Why are you sharing or disseminating these data and what information is necessary to achieve that end?

Let us look at the example of the release of microdata from the 2001 UK census. A survey of users and publications identified that the highly-detailed industry variable in the 1991 census microdata had been used only occasionally whereas the less detailed ethnicity variable was heavily in demand, with users wanting more detail. In 2001 the industry variable was reduced massively in detail and the ethnicity variable was increased from 10 to 16 categories. The net effect was a reduction in measurable disclosure risk (as the reduction in risk arising from the loss of detail on the industry variable outweighed the increased risk arising from the increased detail on the ethnicity variable) but an increase in effective utility (as many users benefited from the increase in detail on ethnicity and fewer suffered from the loss of detail on industry). Things rarely work out that neatly and available resources may restrict one's capacity to carry out anything as extensive as the user survey in this example but nevertheless carrying out a data user needs analysis is an important component of all anonymisation processing.

A second important trade-off is a three-way balancing of data environment risk (risk associated with issues like security, the number of users, governance, etc.), disclosiveness (the properties of the data, given the environment, which make it possible or not to re-identify somebody) and the sensitivity of the data. As is hopefully clear by now, total risk in a data situation is a function of all three of these so that if one increases then the others must be decreased to compensate (if one is to maintain risk at the same functional level). So, for example, if you are comparing a dataset containing mundane information with a second containing sensitive health information, then the environmental and disclosure controls on the latter should, all things being equal, be stronger than on the former. Or if one is thinking of releasing a dataset that was previously only available under special licence as open data, then one must increase the disclosure control applied to the data and/or decrease the sensitivity of the data (by, for example, removing sensitive variables).

This is just common sense but it does suggest a useful insight. In a dynamic data situation, if the data in the original environment are regarded as sufficiently safe (this will normally be so) and if overall risk, taking those three components into account, in the destination data situation is no higher than in the origin data situation, then the destination data situation can also be regarded as safe. This conceptualisation is called *comparative data situation analysis* and is particularly useful for data sharing.

Comparative analysis can also be useful if a *gold standard dataset* exists which has been shared or released in a similar manner to your intended release or share without problems. One such gold standard dataset is census microdata, the record level datasets released from a population census, an example of which we discussed above. In the UK, samples of census microdata have been released under end user license since 1991. As part of the preparation work for these releases, extensive and detailed work on the disclosure control for these datasets is carried out for several years before and after each census. International experts are consulted. Rich and deep technical analyses are conducted. Penetration tests are carried out. To date, there have been no re-identification issues (that anyone is aware of). So, if one has a comparable data situation to that of the release under licence of these microdata, one has an available comparison to a tried and tested data situation and can glean insight from the intensive work that was done. If the risk levels in your data situation are no higher than that, then one can be reasonably confident that they are safe enough.

2.5.2 Data-focused solutions

Data-focused anonymisation solutions require that the data to be released or shared are altered in some way. Usually key variables are removed, obscured or aggregated. Sometimes the same thing is done to those target variables likely to tempt an intruder in order to reduce their sensitivity. We divide these solutions into two types of control: *metadata-level controls* (sometimes called 'non-perturbative methods' or 'non-perturbative masking') where the overall structure of the data is changed and *data-distortion controls* (sometimes called 'perturbative methods') where the data are changed at the level of individual values for individual cases. We will discuss each in turn.

Metadata-level controls

Controls at the metadata level work with the overall structure of the data. The key components of such controls are the sampling fraction, choice of variables, and the level of detail of those variables. In many ways these are the key tools for carrying out practical anonymisation; they are simple to understand and use, do not distort the data and are transparent in their effects.

Sampling

For surveys, the sample fraction is specified by the study design and so its choice often rests outside disclosure control. However for other forms of data there is some value in considering sampling. It cuts down the risk associated with response knowledge by creating uncertainty that a particular population unit is actually represented in the data, so increasing the probability of false positive matches. Even a 95% random sample creates uncertainty and hardly makes a dent in the analytical power of the data.[48]

Impact on risk: Sampling is one of the most powerful tools in the toolbox. The key point is that it creates uncertainty that any given population unit is even in the data at all.

[48] You might wonder what level of sampling fraction is sufficient to impact effectively on response knowledge. There is no absolute firm line, because it will partly depend on other elements in the data situation. However, we have never encountered a use of more than 95% samples and in some (more open) data situations the sampling faction would probably need to be under 50% in order to be effective.

Impact on utility: The impact of sampling is modest; essentially it will increase the variances of any estimates and reduce statistical power. However, if a user wants to analyse small sub-populations the sampling may reduce their capacity to do this.

Choice of variables

An obvious mechanism of disclosure control is excluding certain variables from the released dataset. The data controller can (i) reduce the number of key variables to which a plausible data intruder is likely to have access, or (ii) reduce the number of target variables. These choices flow naturally from the scenario analyses described in Section 2.3.1. With microdata, the choice is whether a variable appears in a dataset or not. With aggregate data, the choices are about which variables will be included in each table. For point-to-point data shares the variable selection will be driven by the requirements of the user although in practice these may be more negotiable than might initially be apparent.

Impact on risk: The impact of variable selection on risk very much depends on the variables selected. If key variables are de-selected the re-identification risk will be reduced. The effect here is to reduce what Elliot and Dale (1999) call *key power*; the capacity of a set of key variables to discriminate between records and produce both sample and population uniques. If target variables are de-selected the sensitivity of the data is lessened and the potential impact of any breach reduced.

Impact on utility: If a variable is critical to a user's analytical requirements then de-selecting that variable will obviously disable the analysis. With releases one is considering how widespread the use is likely to be and whether the goals of release can be met through a more modest variable selection.

Level of detail

Decisions over level of detail complement those over choice of variables. Here you should consider categories with small counts and determine whether merging them with other categories would significantly lower disclosure risk with minimal impact on the informational value of the data. Not surprisingly, many data users would like the maximum level of detail possible on every dataset. But some variables, especially geography and time, can be particularly problematic. Area of residence is a highly visible component of an individual's identity, and so geographical detail is often constrained and data are released at coarser detail than users would like. Similarly, time-based variables, such as exact date of birth, can be straightforwardly identifying when combined with other variables.

Impact on risk: The effect of changing the detail on variables is similar to that of de-selecting variables. It is mainly a mechanism for reducing key power. If a variable has some categories that might be considered sensitive then sensitivity can be reduced by merging these with other categories.

Impact on utility: The impact on utility is similar but more subtle than the impact of removing whole variables. Some variables can be more important than others. Purdam and Elliot (2007) carried out a survey of users to establish the impacts on their analyses of such measures. On most obvious aggregations there was some loss of utility, users reporting that the analysis that they had carried out on the data would no longer be possible.

Distorting the data

The main alternative to metadata controls are various forms of data distortion, which we call *perturbation*. These techniques manipulate the data in order to foil re-identification/subtraction strategies so that an intruder cannot be certain that any match in a re-identification attack is correct or that any zero recovered through subtraction attack is a real zero. In this section, we will look at methods of perturbation that are commonly used for disclosure control.

Data swapping

Data swapping involves moving data between records in a microdata set. A particular form of this, often called 'record swapping', involves swapping the geographical codes of two records.

Impact on risk: Data swapping like most data-focused controls increases uncertainty. However, as Elliot (2000) showed, the impact on general risk measures is quite modest. It comes into its own in situations where multiple data products are being released from a single data source. For example, a sample of microdata with coarse geography (level 1) and aggregate population tables of counts for fine geography (level 2) is a common set of census outputs. Modest data-swapping between level 2 areas within the level 1 areas means the microdata itself is unperturbed. However, the perturbation in the aggregate data will reduce the risk of subtraction attacks and make any attempt to link the fine geography.

Impact on utility: Even done well, the impact on data utility can be significant and it will often affect relationships between variables in an arbitrary and unpredictable manner. For this reason, it is not used routinely in data situations where a single data product is involved.

Overimputation

Overimputation involves replacing real values with ones that have been generated through a model. In order for this to work without badly distorting the data, it may be necessary to allow the original values to be modelled back in. A critical decision when overimputing will be what you tell the user. There are numerous options. You can choose whether to tell them that the data has been overimputed, and if you do then you can also choose whether or not to tell them how many values have been imputed, the model that has been used to do that imputation or even the actual values that have been imputed.

Overimputation can be attractive if you are already using imputation to deal with missing values.

Impact on risk: It is difficult to generalise about the risk impact of overimputation as it depends on the mechanism that is used to decide on the new value, how transparent you are about what you have done and how much overimputation you have done.

Impact on utility: This really depends on how good a model you have used to produce the over imputed values.

Rounding

Rounding is a technique most commonly used with tables of counts. In the simplest form all the counts are rounded to the nearest multiple of a base (often three, five, or ten). Counts which are a multiple of the base number remain unchanged. Normally, the margins are rounded according to the same method of the internal cells. Therefore, in many cases this method does not yield an additive table.[49]

One method of *de facto* rounding which also has some presentational advantages is to release tables of percentages rather than actual counts. Take for example Table 2.8. Looking at this table, we immediately know that any black person living in Anytown earns less than £20 per hour.

[49] An additive table is simply one where the row, column and grand totals are correct. When one rounds the values in a table that may well cease to be true.

	Pay per hour (£ sterling)					
Ethnic Group	**<8.00**	**8.00-9.99**	**10.00-14.99**	**15.00-20.00**	**>20.00**	**Total**
White	10021	19981	49504	38769	1987	120262
Black	1012	876	466	381	0	2735
Asian	1115	1781	1465	1235	116	5712
Other	200	286	134	83	66	769
Total	12348	22924	51569	40468	2169	129478

Table 2.8: A fictitious table of counts showing the pay per hour for residents of Anytown broken down by ethnic group.

Compare this with table 2.9 which presents the same information expressed in terms of row percentages. There are two points here. First, we can no longer tell that the number of black people earning >£20 is zero. In fact the range of possible values here is anywhere up to 16. Second the impact of presenting the table this way is minimal, in terms of what might be considered the underlying message of the data about the wage differential.

	Pay per hour (£ sterling) - n=129478					
Ethnic Group	**<8.00**	**8.00-9.99**	**10.00-14.99**	**15.00-20.00**	**>20.00**	**Total**
White	8%	17%	41%	32%	2%	93%
Black	37%	32%	17%	14%	0%	2%
Asian	20%	31%	26%	22%	2%	4%
Other	26%	37%	17%	11%	9%	1%
Total	100%	100%	100%	100%	100%	100%

Table 2.9: A fictitious table of counts showing the banded pay per hour for residents of Anytown expressed as percentage of the total number of residents of 4 ethnic groups.

Impact on risk: Rounding can be very effective in reducing risks when considering individual tables of counts. Smith and Elliot (2008) demonstrate this with data from the UK neighbourhood statistics. Care must be taken to consider the interactions between multiple outputs and particularly what you are doing about the issue of additivity and consistency between marginal totals in different tables.

Impact on utility: For many purposes rounded frequencies are sufficient and using percentages as a form of rounding can be an even more digestible way of presenting information.

Cell suppression

Cell suppression is a statistical disclosure control technique that can be implemented in various forms whereby the data are only partially released. In one sense, releases of aggregate data are themselves primary examples of suppression, since they are

partial releases of the underlying microdata (or what is sometimes called 'the full table'). If I release two one-way frequency tables, but not the combined table then I am, in effect, suppressing the cross-classification of those two variables. Cell suppression is effectively a more targeted form of this.

Take Table 2.8 again. One alternative is to release Table 2.10 (where the Xs denote the suppressed cells). Note that we cannot simply suppress the disclosive cell (black, >20) as simple arithmetic would allow an intruder to recover it so we must also make what are called complementary suppressions. Another possible suppression pattern is shown in Table 2.11.

	Pay per hour (£ sterling)					
Ethnic Group	**<8.00**	**8.00-9.99**	**10.00-14.99**	**15.00-20.00**	**>20.00**	**Total**
White	10021	19981	49504	38769	1987	120262
Black	1012	876	466	381	66	2735
Asian	1115	1781	1465	1235	116	5712
Other	200	286	134	149	66	769
Total	12348	22924	51569	40468	2169	129478

Table 2.10: A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by ethnic group with cells suppressed in order to reduce disclosure risk.

	Pay per hour (£ sterling)					
Ethnic Group	**<8.00**	**8.00-9.99**	**10.00-14.99**	**15.00-20.00**	**>20.00**	**Total**
White	10021	19981	49504	38769	1987	120262
Black	1012	876	466	381	X	X
Asian	1115	1781	1465	1235	116	5712
Other	200	286	134	83	66	769
Total	12348	22924	51569	40468	X	X

Table 2.11: A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by ethnic group with cells suppressed in order to reduce disclosure risk.

A key point here is that in 2.10 both the users and the intruder can still make inferences about the contents of the suppressed cells. This is not the case in 2.11 because the grand total is suppressed. On the other hand the grand total may well be a piece of information that is published elsewhere and if so it would be simple to unpick the suppressions. The only way to prevent that would be to ensure that the grand total is never published anywhere which may be both impractical and undesirable. For that reason, the pattern in Table 2.10 will generally be preferable.

Why do we say that we can still make inferences about the suppressed cells in Table 2.10? Well, for each of the suppressed cells the value is *bounded* by the other information in the table. Put simply, for each cell there is a limited range of possible values – referred to as *bounds*. The bounds for Table 2.10 can be seen in Table 2.12.[50]

	Pay per hour (£ sterling)					
Ethnic Group	**<8.00**	**8.00-9.99**	**10.00-14.99**	**15.00-20.00**	**>20.00**	**Total**
White	10021	19981	49504	38769	1987	120262
Black	1012	876	466	315 - 381	0 - 66	2735
Asian	1115	1781	1465	1235	116	5712
Other	200	286	134	83 - 149	0 - 66	769
Total	12348	22924	51569	40468	2169	129478

Table 2.12: A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by ethnic group showing the bounds for the cells suppressed in Table 2.11.

Impact on risk: Suppression can be effective in hiding disclosive cells. However you should be aware of the actual intervals that are being implicitly published. As with rounding, care also needs to be taken when releasing multiple tables as it may be possible to unpick the suppressions even if that is not possible when considering each table on its own.

Impact on utility: Users tend to strongly dislike cell suppression. Working with tables with suppressed cells is harder work than say the same tables with rounded values.

Value Suppression

Suppression can also be used for microdata where particular variables can be suppressed for particular cases. For example if you had a 16 year old widower with a child on your dataset you might suppress the age on that case – mark it as missing data in effect. This is an alternative to, and arguably more transparent than, overimputation.

K-anonymisation

K-anonymisation was developed by Samrati and Sweeney (1998) and is a hybrid disclosure risk assessment and disclosure control technique. In essence, it defines a measure of *safe data*:

[50] The example shown here is relatively straightforward. However, precise bounds calculations can be quite complicated. See Dobra and Fienberg (2000, 2001) and Smith and Elliot (2008) for a discussion of the methods required do this.

> A dataset is regarded as k-anonymised if – on all sets of key variables – each combination of possible values of those variables has at least k records that have that combination of values.

This is relatively easy to understand and to implement. There are available open software tools that can semi automate the process.[51] However, its simplicity can be beguiling and the user should be aware that there is no method inherent to the k-anonymity model for identifying either the 'correct' level of *k* or the combinations of the variables that should be considered. Both of these require an understanding of the data environment. Without such understanding, the context is not represented, and the sufficiency of the anonymisation can only be estimated from the properties of the data themselves, which as we have argued misses the point. It also implicitly assumes that either you have a population file or that the intruder has response knowledge (otherwise the k is simply a sample k which could be very misleading and may lead to over-aggregation[52]) and there is no easy way of adjusting the method to deviate from those assumptions.

Another issue with k-anonymity is that it does not allow for attribution disclosure. So if a record shares key attributes with k-1 other data units, that may not help you if all k units share a value on some sensitive attribute. So in Table 2.13 we have k-anonymised the combination of age and sex to k=3 (by in this case merging the 36-64 and 65+ categories). Unfortunately, because all males in the 36+ group have cancer, I can still infer that any 36+ year old male has cancer.

Person number	Age group	Sex	Has cancer
1	0-16	Male	No
2	0-16	Female	No
3	0-16	Female	Yes
4	0-16	Male	No
5	0-16	Female	No
6	0-16	Male	Yes
7	17-35	Male	No
8	17-35	Female	Yes

[51] See for example ARX http://arx.deidentifier.org/downloads/ (accessed 19/3/2016) or µ-ARGUS http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cmu.htm (accessed 19/3/2016)

[52] The degree of aggregation required to achieve the desired level of k will become more severe as the number of data units decreases. So if one is only focused on the data (and not the underlying population) then a sample dataset would be more heavily aggregated than the equivalent population dataset which is clearly counterintuitive unless you are assuming response knowledge.

Person number	Age group	Sex	Has cancer
9	17-35	Male	No
10	17-35	Male	Yes
11	17-35	Female	No
12	17-35	Female	No
13	36+	Male	Yes
14	36+	Female	No
15	36+	Female	No
16	36+	Male	Yes
17	36+	Female	No
18	36+	Female	Yes
19	36+	Female	No
20	36+	Male	Yes

Table 2.13 Hypothetical population microdata for the people living in Anyvillage

To deal with this problem the concept of *l*-diversity was introduced which imposes a further constraint where each equivalence class (group of data units sharing the same attributes) must have multiple values on any variable that is defined as sensitive (or in our terms a target variable). Unlike *k*-anonymity there are various different definitions of *l*-diversity. The simplest is that there has to be at least *l* different values for each sensitive variable within each equivalence class on the key variables.

But it too can lead to counterintuitive outcomes. So in Table 2.14 we have achieved *l*-diversity of 2 but arguably this is a more problematic table rather than less partly because we now know more precisely the type of cancer that the 65+ men have, and partly because that is true for anyone who we happen to know has cancer but do not know which type.

Person number	Age group	Sex	Cancer type
1	0-16	Male	N/A
2	0-16	Female	N/A
3	0-16	Female	Leukaemia
4	0-16	Male	N/A
5	0-16	Female	N/A
6	0-16	Male	Bone Marrow
7	17-35	Male	N/A
8	17-35	Female	Breast
9	17-35	Male	N/A
10	17-35	Male	Leukaemia
11	17-35	Female	N/A

Person number	Age group	Sex	Cancer type
12	17-35	Female	N/A
13	36+	Male	Liver
14	36+	Female	N/A
15	36+	Female	N/A
16	36+	Male	Prostate
17	36+	Female	N/A
18	36+	Female	Breast
19	36+	Female	N/A
20	36+	Male	Prostate

Table 2.14 Hypothetical population microdata for the people living in Anyvillage

To deal with this and other problems with *l*-diversity a third notion, *t*-closeness, has been introduced. This states that the distribution-sensitive variables within each equivalence class should be no further than the threshold *t* from the distribution across the whole dataset.

It would be reasonable to say at this stage that we have moved some distance away from the neat and simple idea of *k*-anonymity. Even if you are using a software package to do the heavy lifting for you, you are still going to need to understand what *k*, *l*, and *t* actually mean for your data and how this relates to what the intruder might be able to do. The risk here is that you make arbitrary decisions led by the privacy model rather than the data situation. We are not averse to the use of privacy models. If used carefully with full awareness of the meaning of the data, *k*-anonymity and its companion concepts can be useful tools in some data situations. However, they are not magic bullets, being neither necessary nor sufficient.

2.5.3 Environment-based solutions

Environment-based solutions essentially control data users' interactions with the data in some way to reduce the degrees of freedom of one or more (and usually all four) of the elements of the data environment (other data, agents, governance processes and infrastructure). The key point to remember is that one cannot make a judgement about whether data are anonymised or not without reference to their environment. The implication of this is that by the operation of environmental controls one can anonymise the data just as effectively as through controls on the data themselves.

Following Duncan et al (2011), environmental controls can be broadly characterized as answering 'who', 'what', and 'where and how' questions:

- **Who** has access to the data?
- **What** analyses may or may not be conducted?
- **Where** is the data access/analysis to be carried out **and how** is access obtained?

These questions are interrelated – a decision about one has implications for the others. We will consider each in turn.

Who can have access?

The 'who' question is essentially all about *agent control*. In crude quantitative terms the risk level of 10 people accessing your data is considerably smaller than if you open it up to 10,000. Beyond the simple additive effect of more people contributing some quantum of risk there is the additional effect that opening up access necessarily implies more relaxed governance. With 10 people it is possible to think about vetting procedures, but vetting 10,000 will inevitably be less sensitive and more routine.

This raises the question of how the data controller identifies those classed as 'safe people'; is there a method that establishes those individuals or organisations that the data controller should trust and those that he or she should not? At a high level, organisations or individuals with a track record of good practice in data security and stewardship may be given greater data access rights than those without. Often, restricted access conditions stipulate that users must have specified credentials to get access to data. Here are some criteria by which a data controller might assess a potential data user.

1. Whether he/she is associated with some organisation that can assure compliance with the data controller's data access requirements.
2. If the proposed use is for research then the researcher is able to demonstrate the ability to do research of scientific merit.[53]
3. Whether he/she has undergone some sort of 'safe user' training. In the UK a consortium of research data centres has recently developed nationally-based training for the certification of researchers. Attendance at this training and passing a subsequent test is a requirement for data access.[54]

---

[53] It may not seem immediately obvious why this is in the list. However, remember that we are in the game of risk management and there has to be some benefit to counterbalance the risk. Sharing data (and therefore taking a risk) for a piece of work with no value would not meet this requirement.

[54] The UK Statistics and Registration Services Act (2007) defines the notion of the 'approved researcher' and approved projects which – as implemented – covers much of this ground. Similarly,

As in all anonymisation matters the key is proportionality. The degree to which *agent controls* should be applied will be related to the disclosiveness and sensitivity of the data and inversely to the degree of other environmental controls.

What analysis is permitted?

Governance control can constrain the projects that can be undertaken with the data. This may be in the form of categorically prohibiting certain types of analysis or may require a project approvals process. For example, the UK Administrative Data Research Network[55] has a formal project approvals panel. Potential users of the service have to convince the panel that the project has scientific merit, is feasible, will have public benefit and have a negligible impact on privacy.

A related restriction is controlling the output. In a strongly controlled environment, some sort of output control is usually necessary. The intuition here is that if the data themselves are disclosive (in an open environment) then analytical outputs will also have the potential to be disclosive. Outputs are after all simply a form of data and so, as we will discuss in chapter 2, the publication of the results of analyses (which is usually what is intended) creates a dynamic data situation.

In all output checking, what is in essence being checked is whether it would be possible to recover (some of) the underlying data from the output. As a simple example, with risky tabular frequency data one typically would not permit unrestricted requests of multivariate tables of counts,[56] since a sequence of such requests can be used to recover the original data. So, if a user were to request such cross-tabulations then the request would have to be denied. However, the problem goes beyond this situation. For example, using any regression model in combination with residual plots, it is possible to recover some of the original data used to generate the model. At this point, developing valid output checking processes that could be automated is an open research question. Therefore, output needs to be checked manually by data centre staff with some expertise.[57]

---

data resources such as the UK's ADRN (see below) invariably have policies covering issues such as accredited users, feasibility of projects, breaches policy and procedures etc. See for example, https://adrn.ac.uk/using-the-network/documentation.

[55] www.adrn.ac.uk [accessed 30/5/16]

[56] These are cross-tabulations of two or more variables.

[57] Recently, progress has been made with automating at least some of the functionality of output checking (see for example Thompson et al (2013) and O'Keefe et al (2013)). However, it would be fair

One way to reduce the burden on the output checkers, used by, for example, ONS's virtual microdata laboratory[58] is to define a very conservative class of outputs as 'safe' and then leave it to the user to demonstrate that anything not on the list is also safe.

Where and how can access be obtained?

In many ways the where and how questions are the key drivers in determining the type of environment that you are working in. There are four modes of access that are currently used for disseminating data for use outside of organisational boundaries:

1. Open access
2. Delivered access
3. On-site safe settings
4. Virtual access

Open access

Open access (or what can be called *unrestricted access*) has always been used for publishing some census tabulations and headline administrative data. An instance of free access is the UK's Neighbourhood Statistics Service (NSS).[59] Neighbourhood Statistics are intended as public use data. NSS imposes no restrictions on who can access the data, or on what they can do with them. Also, there is usually no monitoring of users or what they are doing. Another mechanism by which data can become open is a freedom of information request. These requests are in effect a request to make the data open and web sites – such as My Society's www.whatdotheyknow.com – ensure that FOI requests are published.

Until 25 years ago, the dissemination medium of such data was paper-based, usually in the form of thick volumes of tables. However, web delivery is now far more common, and this has opened up datasets for much wider use. In the UK there is significant pressure arising from both demand and government policy to make more government data available openly. This has led to the development of the Open Government Licence[60] by The National Archives. This licence specifically does not apply to personal data/information. The point of this is to underline that open data

---

to say that this is work in progress and by no means does it give full coverage of all possible types of output.

[58] See Office for National Statistics (2016) for more detail.

[59] The UK's Neighbourhood Statistics Service http://www.data4nr.net (accessed 22/5/15) provides local area indicators derived from administrative records of multiple government agencies.

[60] See National Archive (2016)

environments are really only appropriate to data that are either *apersonal* or have been through an extremely robust anonymisation process that ensures with a very high degree of confidence that no individual could be re-identified and no statistical disclosure could happen.

Delivered access

Delivered access is a more restricted form of access, in which access to the data is applied for and the data are delivered to the user, most commonly through an Internet portal or possibly via encrypted email. The former is common in cases where delivery is potentially to a community of many users (the UK Data Archive[61] is an example of this). The latter is perhaps more common where the data situation is a single site-to-site share. It is important not to forget that the transfer medium is itself an environment, and that therefore one needs to model potential media as data environments as well in order to decide the appropriate means of transfer.

Usually, as in the example of the Data Archive, the process of applying for a copy of the data requires the user to specify what they are to be used for and invariably he or she is required to agree to specified conditions on a licence for data access. We discuss such licences below.

On-site safe settings

On-site safe settings are regarded as the strongest form of restricted access, usually including a high level of *security infrastructure control*. The data user applies for access to the data in a particular location — often in the offices of the data controller or otherwise at a research data centre (RDC) that has been established by the data controller.[62] Often, the users are required to analyse the data on a dedicated standalone computer and are restricted in the software that they may use. There are also often numerous *governance controls* in place. For example the user may:

1. Not be permitted to take in data transport devices such as USB drives or mobile phones.
2. Not be allowed to copy down anything that appears on the screen.
3. Be required to log in and out of the facility.
4. Only attend at pre-booked days and times.

[61] http://www.data-archive.ac.uk/ (accessed 30/5/16).

[62] Examples are the Administrative Data Research Centres in the UK (www.adrn.ac.uk), the US Federal Statistical Research Data Centres (http://www.census.gov/about/adrm/fsrdc/locations.html) and the Research Data Centre (FDZ) of the German Federal Employment Agency (http://fdz.iab.de/en.aspx).

5. Be required to sign a user agreement stipulating that they will adhere to conditions of access such as those specified in (1)-(4) above and undertake not to attempt to identify any individuals from a de-identified dataset.

The user will be allowed to take away some analytical output, but only after it has been checked by output checkers for disclosiveness.

On-site safe settings may be considered as less than ideal by researchers because (1) travel to one of these sites is expensive, (2) the facility is only open at certain hours, (3) computing facilities may be unfamiliar or inadequate, (4) Internet access may not be available, and (5) it requires users to work in unfamiliar surroundings. However it is worth remembering that such arrangements facilitate research on vitally important but sensitive topic areas that might not be possible in other types of settings.

An alternative approach (used for example by the ONS Longitudinal Study)[63] is that the researcher submits the syntax for their analysis software to a dedicated unit which – if it approves it – then runs it. However, this rather awkward approach is being superseded by virtual access systems.

Virtual access

Virtual access is now widely regarded as the future of research data access. It combines many of the advantages of the physical safe setting with much of the flexibility of having a copy of the data on one's desktop.[64] There are two variants on the virtual access theme: *direct access* and *analysis servers*.

Direct virtual access uses virtual remote network-type interfaces to allow users to view, interrogate, manipulate and analyse the data as if it was on their own machine. There are two critical differences between direct virtual access and delivered access. Firstly, output is typically checked in the same manner as in an on-site safe setting. Secondly, there is still no possibility of a user directly linking the accessed dataset to another dataset (because dataset uploads are not possible) and this restricts the number and type of disclosure scenarios that the data controller needs to consider.

---

[63] See Office for National Statistics (2016b).

[64] An intermediate hybrid approach is where safe rooms or 'pods' are installed at user institutions as a semi controlled medium for virtual access. So the user will have to go to the local safe room, but this will involve minimal travel and is therefore less restrictive than an on-site lab. This is being explored as a method for allowing researchers to have access to administrative data within the UK.

Analysis servers go one step further in not allowing direct access to a dataset while allowing the user to interrogate it. In such systems data can be analysed but not viewed. Usually, there is a mechanism for delivering the analysis (for example through uploading syntax files for common statistical packages or, occasionally, through a bespoke interface). The analysis server will return the results of the request for analysis, usually after they have been checked for disclosiveness. From the data controller's viewpoint, the advantages over direct virtual access are twofold: (i) because the user cannot see the data the risk of spontaneous recognition of a data unit is all but removed, and (ii) there is no risk of the screen being seen by somebody who is not licensed to use the data. The disadvantage from the user's point of view is that it is more difficult to explore the data.[65]

Licensing

Another *governance control* tool a data controller has is licensing, often used in conjunction with other restricted access mechanisms. Licences can be used as a pro forma to be signed by a set of users or in a bespoke data sharing agreement for site-to-site shares. Some common themes in such licensing agreements are:

1. Specification of those permitted access (*agent controls*).
2. Data security requirements (*infrastructure controls*).
3. Restrictions on use, particularly prohibition against linking with other files and on deliberate re-identification (*other data and governance controls*).
4. Requirement to destroy the data once the use is complete (*governance controls*).

The function of licensing is threefold:

1. It clearly distinguishes between those individuals or organisations the data controller trusts and those that it does not.
2. It is a framework for specifying the conditions under which access can occur.
3. It can specify sanctions or penalties should the individual/organisation transgress on those access conditions.

It is possible to have licensing at graded levels, with different users having access to data with different levels of disclosure risk (and therefore presumably different levels of data utility). In the UK, the ONS currently makes a distinction between public, research and special licence levels of access. In such a regime, an inexperienced researcher might be subject to stricter conditions than a professor of

[65] For a useful discussion of the different types of virtual access systems and why a data controller might choose one over another see O'Keefe et al (2014).

long standing. So, as a general mechanism for the dissemination of data for research purposes it might be criticized on fairness grounds.

By including some *infrastructure and governance controls* the licence allows the data controller to maintain some control over the security of the data and can also provide guidance to the data user regarding good practice. If the data are being provided to the user at their site then various physical and computer security conditions might be required. Here is an example of a set of requirements that might be included in a licence for a single site-to site-share:

1. Data must be stored in a dedicated secure data lab.
2. There must be an independent locking system (unmastered) to the data storage area.
3. There must be extra security at all possible primary and secondary points of entry, extra locks on doors, bars on windows, etc.
4. Data must be stored on a standalone machine.
5. Multiple passwords are required to access the data.
6. Devices such as external disc drives/USB ports must be disabled.
7. Output must not be removed from the data lab and must be destroyed when finished with.
8. Entry to the data lab must be limited to particular staff.
9. Log books must be kept of access.

As well as providing actual security, imposing such conditions may also be intended to change the mind-set of the user, who will hopefully react to them by being more security-aware. The flip side of this is that these conditions may place awkward obstacles in the researcher's usual research method.

Another type of commonly employed licence condition asks the user to agree to restrictions on what they can do with the data – in particular, not linking it with other datasets that contain direct identifiers.

Function 3 of the licensing process involves the sanctions that can be applied to users or their organisations for non-compliance with the licence conditions. In order to serve as deterrents for non-compliance, they must be enforceable. Typical sanctions are fines and removal of the right to access the data. For example, from the Statistics Canada Research Data Centres Program: 'Researchers whose projects are approved will be subject to a security check before being sworn in under the Statistics Act as "deemed employees." Deemed employees are subject to all the conditions and

penalties of regular Statistics Canada employees, including fines and/or imprisonment for breach of confidentiality.'; Statistics Canada (2015).

The threat of sanctions will be taken most seriously if the data user or their organisation is subject to a security audit by the data controller. While an audit can be costly to both the data controller and the user, a licence without such a stipulation may not be taken seriously.

Overall, licensing can be a useful way to decrease disclosure risks for certain uses of a disclosive dataset by researchers, especially when explicit or implicit sanctions can be invoked. Although it is commonplace for users to have to sign access agreements for routine data access requirements, beyond giving the user cause for reflection at the point of access there is little or no enforceability in such agreements. The question of whether agreements that are not directly enforceable have real impact is of course a major question in many areas of society. Answers will vary depending on history, social context, existence of informal controls, etc.

Summary

Environment-based controls do provide the potential to reduce the risk of disclosure significantly, possibly more so than can be achieved for the same utility impact by manipulating the data themselves. All of these controls affect at least one of the four elements of the environment (other data, agents, governance processes and infrastructure) and ultimately disrupt the ability of a (malevolent) user to connect identification data to anonymised data.

2.6 Why ethics is an important issue in Anonymisation

It is not always immediately obvious why ethical considerations have a role to play in the process of anonymisation. Most readers will understand that the processing of personal data is an ethical issue but once data are anonymised are our ethical obligations not dealt with? This is an understandable confusion which arises in part from a conflation of legal and ethical constraints. Legally, functional anonymisation is sufficient but this might not be true ethically. There two primary reasons why we need to consider ethics beyond the law:

1. Data subjects might not want data about them being re-used in general, by specific third parties or for particular purposes.
2. We are not dealing with zero risk.

Before discussing this further, we will place a caveat on what we are about to say. The ethics of privacy, data sharing and data protection are hugely contested and this element of the framework is necessarily the most subjective and pre-theoretical. The reader may well have a different view about what is important, particularly about the thorny issue of consent. However we believe the ideas that we present here are consistent with the general approach we are taking and provide a practical method for incorporating ethical thinking into anonymisation decision making.

There is growing evidence that data subjects are concerned not just about what happens with their personal data but also about the anonymised data derived from their personal data.

On this point Iain Bourne from the ICO notes:

> *We do hear – for example from telecoms companies – that customers are increasingly objecting to their data being used for x y and z even in an anonymised form – and I don't think they draw a personal data/non-personal data distinction and why should they? I predict that this form of consumer objection will become much more of an issue.* (Bourne 2015).

There may be many reasons why data subjects object to the reuse of their data. For example I might be unhappy about my data – even anonymised – being reused by a particular type of organisation (perhaps an extreme political group, arms manufacturer or tobacco company). Perhaps I do not want my data to be reused to make a profit for someone else, or I may be simply unhappy that I have not been asked.

For example, O'Keefe and Connolly note the possibility of moral objections to particular re-use:

> *The use of an individual's health data for research can be viewed as participation by that individual in the research. An individual may have an objection to the purpose of the research on moral grounds even when there is no risk of identification or personal consequences.* (2010: 539).

Or data subjects may object to a data reuse on the grounds that it serves a perceived narrow (self) interest or because it has no clear benefit for them or the wider public. For example the Wellcome Trust (2016) state:

> *Overall, the research showed that most people were extremely wary of insurance and marketing companies using anonymised health data. These companies were seen to be*

acting against the interests of individuals, motivated by their own private interests with little or no public benefit. (2016:2).

Or an objection to data reuse might simply arise because the data subject gave their data for one purpose and you have used it for a different purpose.

In short, there are numerous reasons why data subjects might object to their data being reused. This brings us to the thorny issue of consent. In principle consent is a straightforward idea. You ask the data subjects 'can I do X with your data?' and they say yes or no. However, in practice the situation is much more complicated than this. Firstly, consent is layered. Secondly, the notion of consent is interlaced with the notion of awareness. This produces what we refer to as a scale of information autonomy. Consider the following questions:

1. Are the data subjects aware that their data have been collected in the first place?
2. Have the data subjects consented to the collection of their data?
3. Were the data subjects completely free to give consent to the collection of their data or have they agreed to collection because they want something (a good or service) and are required to hand over some data in order to obtain it?
4. Are the data subjects aware of the original use of their data?
5. Have the data subjects consented to the original use of their data?
6. Have the data subjects consented in general to the sharing of an anonymised version for of their data?
7. Are the data subjects aware of the specific organisations that you are sharing their anonymised data with?
8. Have they consented to your sharing their data with those organisations?
9. Are the data subjects aware of the particular use to which their anonymised data are being put?
10. Have they consented to those uses?

The more 'no's that you receive from the above list, the less autonomy the data subjects have. What does this mean in practice? Put simply, as the data subjects become less autonomous the less able are they to take responsibility for what happens to their data and therefore the greater your own responsibility. We shall see how this plays out in your anonymisation process in component 5 in the next chapter.

Of course the astute reader will have noted that not all (and possibly none) of the questions have straight yes or no answers. Awareness is a nuanced concept. For

example, take question 1; I might be generally aware that I am being caught on CCTV every day but not know about every (or even any) specific instance of that. Or I might be aware that I have been caught but not know what happens to the film next and so on. Similarly I may have *de facto* consented to a particular piece of data processing but not have understood what I have consented to. Am I, in fact not even aware that I have consented? So awareness and consent interact.

What does this complex autonomy soup actually mean? You might be expecting us to say at this point that you should be seeking informed consent if at all possible but we are not going to do that. Given the current state of the information society this is both impractical and undesirable. Obtaining consent of any sort is complex. Obtaining real informed consent would – just as a starting point – require re-educating the whole populace and even then giving consent for every piece of processing for every piece of data is not something that most, if not all, people are going to engage with consistently (if you have never ticked the box to agree to T&Cs on a web site without having first read them, please get in touch with us as we would like to know what that is like). This is not to say that well thought out consent processes do not have their place – they most certainly do – but they are not a panacea.

Ok so what is the point here? It is simply this: if you pose the questions above and the answers are mostly in the negative then your data situation is more sensitive. The notion of a sensitive data situation is key here; it is a connecting concept which enables clearer thinking about ethics and the reuse of (anonymised) data. We will come on to what you need to do about sensitive data situations shortly but is there anything else that heightens sensitivity?

Beyond explicit consent, the question of whether a particular share or release conforms to the data subjects' reasonable *expectations* is also important. Nissenbaum's (2004, 2010) description of privacy is useful here. She describes privacy not as a right to secrecy nor as a right to control 'but a right to appropriate flow of personal information' (2010:127). To help tease out the appropriate flow, and what your stakeholders expectations may be, we draw (loosely) on Nissenbaum's concept of contextual integrity. Contextual integrity is a philosophical approach for understanding privacy expectations in relation to the flow of personal information and can usefully be applied to shed light on why some flows of personal data cause moral outrages. This approach uses the notion of context, roles and data (to be transmitted) as a framing tool for evaluating whether a flow of data is likely to be considered within or outside of (i.e. violating) expectations.

We argue that the principles of the concept 'contextual integrity' can usefully be applied to the flow of anonymised data for the purpose of helping practitioners to make well thought out and ethically sound decisions about how they reuse data.

To untangle this complex notion for practical use you will need to think about the terms of roles and relationship between you and the proposed receiver of your anonymised data, and the purpose of the share/release. The complexity of the questions you will have to ask yourself will depend on the complexity of your data situation. But here is how they might look for a simple site-to-site share of data:

1. Do you (the sending organisation) have a relationship with the data subjects?
2. Does the receiving organisation have a relationship with the data subjects?
3. Do you and the receiving organisation work in different sectors?
4. Is your organisation's area of work one where trust is operationally important (e.g. health or education)?
5. Is there an actual or likely perceived imbalance of benefit arising from the proposed share or release?

Here the more questions you answer yes to, the more sensitive your data situation is.

Finally, the data themselves can have properties that make the data situation more or less sensitive. Three questions capture the main points here:

1. Are some of the variables sensitive?[66]
2. Are the data about a vulnerable population? A vulnerable group is one where its members lack capacity (partial or full) to make informed decisions on their own behalf. Examples of vulnerable groups include children, adults with mental impairments or subpopulations constructed out of a category that would itself be considered sensitive – for example a minority ethnic group or AIDS sufferers.
3. Are the data about a sensitive topic? The topic area might be considered sensitive rather than, or as well as, the variables within the anonymised dataset because, for example, it involves particular public interest issues, or ethically challenging issues.

[66] The UK Data Protection Act (1998) identifies the following as sensitive: (a) The racial or ethnic origin of the data subject, (b) Their political opinions, (c) Their religious beliefs or other beliefs of a similar nature, (d) Whether they are a member of a trade union, (e) Their physical or mental health or conditions, (f) Their sexual life, (g) The commission or alleged commission by them 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. This list is widely regarded as being insufficient and out of date; financial information is, for example, absent. A simple litmus test is: would a reasonable person regard these data as sensitive?

Again here, the more questions you answer yes to, the more sensitive your data situation.

So we have three components of data situation sensitivity: consent, expectations and data. These components interrelate. So trust questions (expectation sensitivity) will be more significant where the data are about a vulnerable population (data sensitivity).

Underlying this notion of sensitivity is one of potential harm. The notion of harm is commonly measured in quantitative/economic terms such as financial loss but it is also recognised that it can be felt in subjective ways such as loss of trust, embarrassment or loss of dignity. Harm might also occur at the individual, organisation or societal level. The latter two might arise because of knock-on consequences of a reuse of data that violates expectations (whether it is formally a privacy breach or not) and leads, for example, to the shutdown of data access and societal benefit not accruing because people become less likely to respond to surveys, provide accurate data etc. You should not underestimate harm at these levels – it means that all organisations who deal with data have a collective interest in everyone getting reuse right.

Harm felt subjectively is recognised in law – e.g. Article 8 of the European Convention of Human Rights stipulates that everyone has the right to respect for his or her private and family life, home and correspondence. Article 12 of the Universal Declaration of Human Rights (1948) goes even further: 'No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.' The concept of 'a right to a private and family life' encompasses the importance of personal dignity and autonomy and the interaction a person has with others, both in private and in public.

Hopefully you can see how the notion of data situation sensitivity allows us to gain traction on the somewhat intangible notion of potential harm and that by asking yourself questions about consent, awareness, expectations and the data, you are able to formulate a practical understanding of the concept. In chapter 3, component 5 we will examine how it is possible to apply this in your own data situation.

2.7 Chapter Summary

In this chapter we have introduced the key concepts that you need to understand the anonymisation decision-making framework. This has covered quite a breath-taking

range of topics from the law and ethics through to statistics and notions of risk and likelihood. This interdisciplinary collection of ideas, concepts and techniques forms the toolbox that you will need to use but hopefully, as you have worked through the chapter, you have picked up on the key take home message: anonymisation is a tractable problem.

The key unifying concept underlying our approach is the data situation; the relationship between some data and its environment. Although this is in itself a complex concept, grasping it will enable you to understand the necessary components of any anonymisation decision making that you have to do. In the next chapter we will use these concepts working through the framework component by component.

Chapter 3: The Anonymisation Decision-Making Framework

3.0 Introduction

In chapter 2 we described the core concepts that underlie the notion of anonymisation. We also established the need for structured guidance for anonymisation decision-making, to enable data custodians to make robust, principled decisions about the data they are responsible for. In this chapter, we describe the *Anonymisation Decision-Making Framework* (ADF), which meets this need.

The ADF is made up of ten components, and we will describe each of these components in detail in this chapter:

1. Describe your data situation
2. Understand your legal responsibilities
3. Know your data
4. Understand the use case
5. Meet your ethical obligations
6. Identify the processes you will need to assess disclosure risk
7. Identify the disclosure control processes that are relevant to your data situation
8. Identify who your stakeholders are and plan how you will communicate
9. Plan what happens next once you have shared or released the data
10. Plan what you will do if things go wrong

These ten components comprise three core anonymisation activities:

- ***A data situation audit*** (components 1-5). This activity will help you to identify and frame those issues relevant to your data situation. You will encapsulate and systematically describe the data, what you are trying to do with them and the issues thereby raised. A well conducted data situation audit is the basis for the next core activity.
- ***Risk analysis and control*** (components 6-7). Here you consider the technical processes that you will need to employ in order to both assess and manage the disclosure risk associated with your data situation.
- ***Impact management*** (components 8-10). Here you consider the measures that should be in place before you share or release data to help you to communicate with key stakeholders, ensure that the risk associated with your

data remains negligible going forward, and work out what you should do in the event of an unintended disclosure or security breach.

How you use the framework is likely to depend on your level of knowledge and skills as well as the role you play in your organisation. Some might use it for knowledge development purposes, to understand how a privacy breach might occur and its possible consequences, or to develop a sound understanding of the important issues in the anonymisation process. Others might use it directly to support their management of the risk of a privacy breach, to reduce it to a negligible level.

Anonymisation is not an exact science and, even using the ADF at this level, you will not be able to avoid the need for complex judgement calls about when data is sufficiently anonymised given your data situation. The ADF will help you in making sound decisions based on best practice, but it is not an algorithm; it is an approach whose value depends on the extent of the knowledge and skills you bring to it. You may still need expert advice on some parts of the anonymisation process, particularly with the more technical risk analysis and control activity. However, even in such a situation the ADF can still be very useful; you and your expert will have more fruitful discussions, make quicker progress and will be more likely to produce a solution that works for you if you properly understand your data situation. Consider the ADF as a member of your team; it will not solve all your problems, but will provide graded support appropriate to your own level of expertise.

A final point before we launch into describing the framework in detail: in all likelihood you will need to adapt the framework to suit your own needs. Whether you use the ADF to expand your knowledge or to support decision-making, it is important to recognise that it is not a simple check list that you can run through in a linear fashion and tick off as you go down the page. All the important considerations are there but you will need to think how they relate to and impact on each other. Some aspects may be more important than others for your data situation. Most importantly, in applying the framework you should keep clear in your mind that the objective is to disseminate *safe useful data*.

3.1. The data situation audit

The data situation audit is essentially a framing tool for understanding the context of your data, and therefore to help scope the anonymisation process appropriately for you to share your data safely. It will help you to clarify the goals of the process

and will enable the more technical aspects of the anonymisation process (components 6 and 7 of the ADF) to be planned and conducted more rigorously.

Component 1: Describe your (intended) data situation

In chapter 1 we introduced the term *data situation* to refer to the relationship between some data and their environment. So for example, your organisation itself will constitute an environment, whilst any proposed share or dissemination would constitute another environment. These environments will have different configurations of the same core features: people, other data, infrastructure and governance structures.

Data situations can be *static* or *dynamic*. A static data situation is where there is no movement of data between environments; a dynamic data situation is where there is such movement. By definition all data shares or dissemination processes take place within dynamic data situations in which data are intentionally moved from one environment to another. A dynamic data situation might be relatively straightforward involving the movement of data from just one environment to another environment. Often though, it is more complex involving multiple environments.[67]

At this stage we want to familiarise you with the idea of data moving between environments. Whilst data environments can be thought of as distinct contexts for data they are interconnected by the movement of data (and people) between them. As we have said previously, by mapping the data flow from the point at which data is collected to the point after which it is shared or released you will be able to define the parameters of your data situation. We will illustrate this idea further using two examples of data flows across environments.

Data situation: simple share

In the first example we look at the data flow across environments involving data that have been subject to anonymisation.

Imagine that PubT (a franchised public transport provider) collects personal data from its customers relating to public transport usage. PubT plans to share an

[67] Actually there are more variations in data situations than this distinction allows. We do not here consider issues arising from multi party computation for example. However for the purposes of exposition we will restrict ourselves to the relatively simple case of a unidirectional sharing/dissemination process.

anonymised version of the data with the Local Authority of Barsetshire which wants to use it to support (better) provision of public transport. PubT anonymises the data by removing the direct identifiers, i.e. the customers' names and addresses, and by aggregating the detail on several key variables. However, it leaves some key variables – which are of particular interest to Barsetshire – unchanged. In this environment PubT is the data controller for this data because it determines the purposes for which, and the manner in which, the data are processed. Call this environment 1.

Barsetshire signs a contract with PubT which (i) enables it to analyse the data (for a purpose other than that for which it was collected), (ii) proscribes it from sharing or releasing any part of the data without the prior agreement of PubT and from holding the data for longer than a specified time period and (iii) requires Barsetshire to keep the data securely and safely destroy it once it has finished using it. After this contract is signed, the anonymised dataset is passed to Barsetshire, so call Barsetshire's arrangements environment 2.

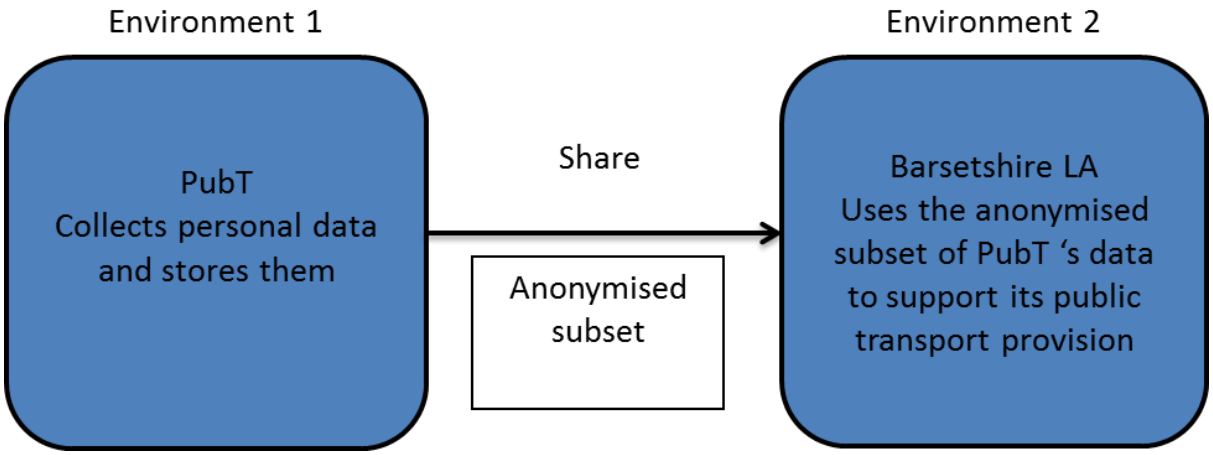


Figure 3.1: Data flow between two environments

Figure 3.1 illustrates the intentional movement of data from environment 1 to environment 2. The data flow between PubT and Barsetshire defines the parameters of the data situation. By using a contract to stipulate how the data can be processed and accessed, PubT is placing controls on governance processes and infrastructure within environment 2 and thereby controls the disclosure risk associated with the data situation. The (anonymised) data within Barsetshire's environment is considered low risk even though it contains some detailed key variables. This is because the environment is restricted – few people have access to the data and their use of the data is clearly defined and managed. Conversely, in this scenario the data would not be considered safe within a less restricted environment, such as an open

access environment, because no such control restrictions would be in operation. This may seem obvious, but failure to understand the basic point that data releases need to be appropriate to their release environment is the primary cause of the well-publicised examples of poorly anonymised datasets such as the AOL[68], Netflix[69] and the New York taxi driver[70] open datasets.

Data situation: simple share with secondary open release

Consider this example further and imagine Barsetshire would like to release part of the data openly. For example, it might want to publish aggregate cross-tabulations of public transport use by key demographics as part of a transparency initiative. Aggregate outputs are still data and so such a release extends and indeed complicates the data situation. The 3rd environment in the chain is the open environment. The new picture of the data flow is shown in Figure 3.2.

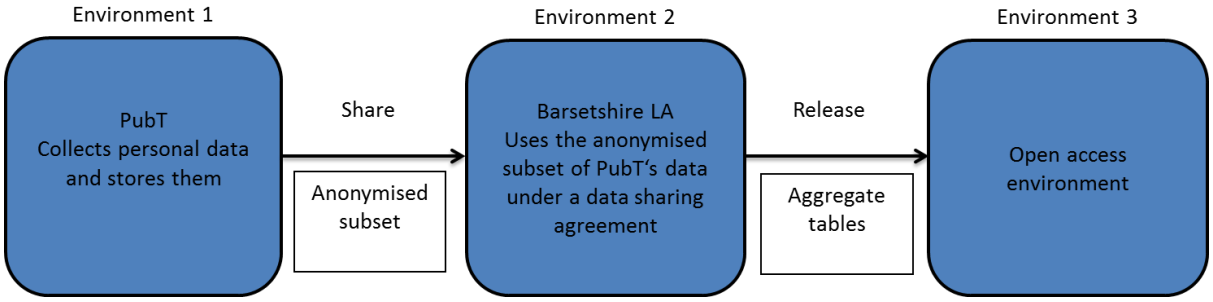


Figure 3.2: Data flow between multiple environments

The data flow between PubT, Barsetshire and the open access environment defines the parameters of Barsetshire's data situation for the anonymised public transport data.

Barsetshire, as stipulated in its contract with PubT, cannot release the anonymised data given to it in its original form without permission from PubT. This is because PubT is the data controller for the (personal) data and as such retains full data protection responsibilities for it (we shall look at this further in component 3 below). Prior to releasing any data, Barsetshire should carry out a disclosure risk audit of the open access environment[71] and then further anonymise the intended disseminated

---

[68] See Arrington (2006) http://tinyurl.com/AOL-SEARCH-BREACH

[69] See CNN Money (2010) http://tinyurl.com/CNN-BREACHES

[70] See Atokar (2014) http://tinyurl.com/NYC-TAXI-BREACH

[71] In component 6 of the ADF we set out how you can go about assessing the risk associated with the open data environment. This includes examining what other data sources might be available and sketching out the 'who', 'why' and 'how' of a potential statistical disclosure.

data product as necessary given the likely use case(s) (this is covered in component 4 of the ADF).

Data situation: simple share with secondary controlled release

In this example we look at the data flow across environments involving personal and de-identified data. Imagine that Barsetshire collects public health data for its area. It has powers under statutory law to share (some of) the public health data with the Department of Social Affairs (DoSA) to support its work on health promotion and disease prevention. The data share is formalised under a data sharing agreement (DSA)[72] which stipulates both Barsetshire and DoSA as data controllers in common for those data. This means both organisations have full data protection responsibilities. We call Barsetshire council environment A.

The DoSA as part of its remit for health promotion (and in accordance with its agreement with Barsetshire) creates a de-identified subset of the data and makes it available within a secure setting for reuse by approved accredited researchers. The DoSA is environment B.

The secure setting is designed in such a way as to ensure that the de-identified data are functionally anonymous. It places restrictions on who can access the data, on how they can be accessed, and on what auxiliary information can be brought in and out of the secure lab environment. The secure lab is environment C.

An approved accredited researcher carries out her data analysis in the secure lab producing statistical output, such as regression models, that she will need to write up for her research. These outputs are first checked by secure lab staff to ensure that they are not disclosive, so they are passed as 'safe'. The researcher duly writes up and openly publishes her research, which contains some of the analytical output. The publication of the research is a fourth environment, which we call environment D.

[72] A data sharing agreement should set out a common set of rules to be adopted by the organisations involved in the share. It should cover such issues as: (i) the purpose(s) of the share; (ii) the recipients of the share and the circumstances in which they will have access; (iii) the data to be shared; (iv) data security; (v) retention of shared data; (vi) sanctions for failure to comply with the agreement. For further information see UK: Information Commissioner's Office (2011).

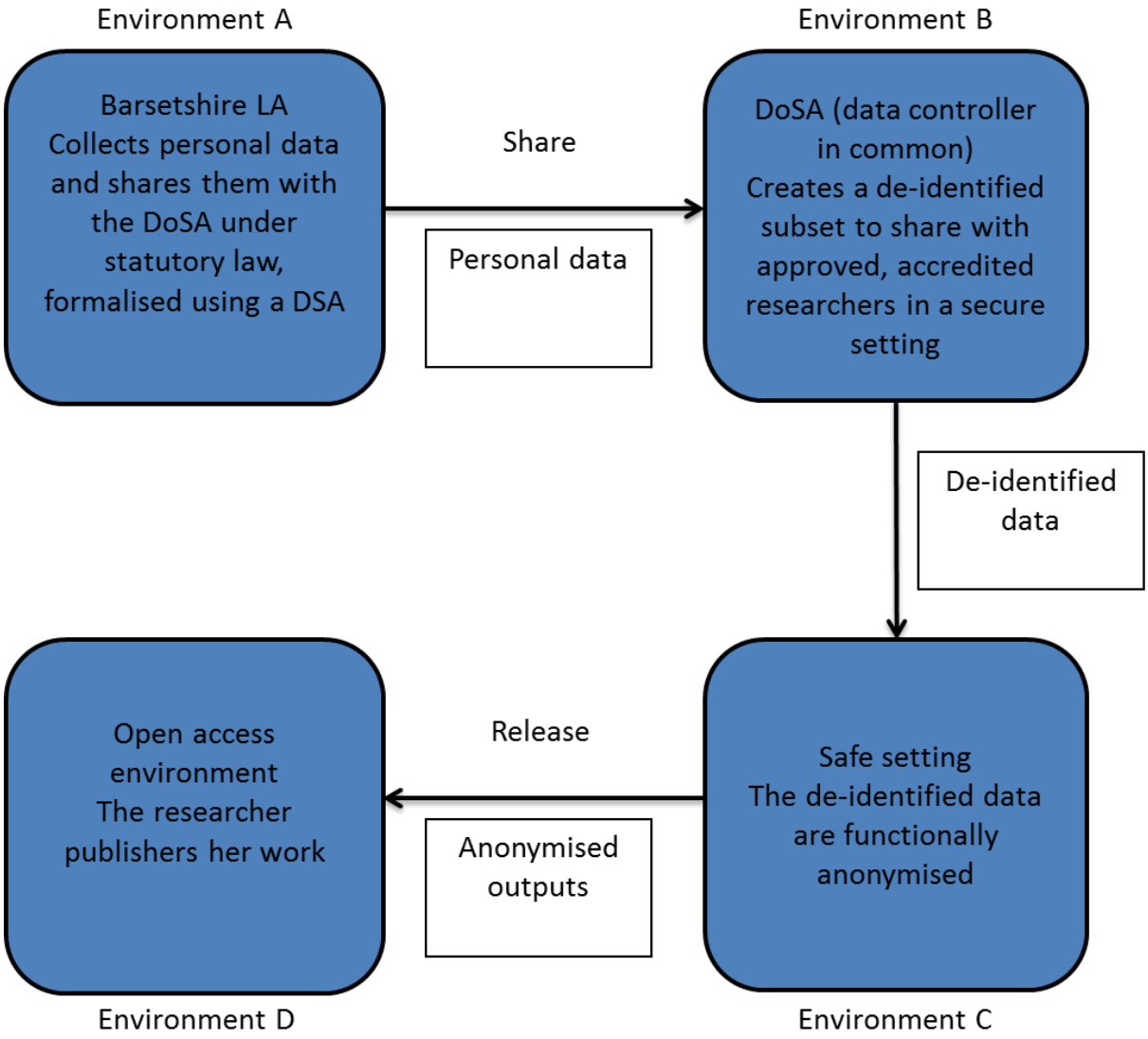


Figure 3.3: Movement of data across data environments

As in the first example, one of the key issues which we particularly wish to highlight is that data in one environment may be considered sufficiently anonymised (for example the de-identified data in the secure setting), but in a different environment (such as the researcher's publication) this may no longer be the case at all. Hence in this example the researcher's analytical outputs have to be checked and verified as 'safe' by the secure lab before she can take the data away with her.

It is worth stressing that whilst a data situation might be complex, it should not be considered a problem so intractable that you feel it safer not to *even consider* sharing or releasing your data. Of course, that might be the conclusion that you come to after you have worked through the ADF but it should not be the starting position. You should not lose sight of the enormous range of benefits that can and do come from sharing and opening data.

Component 2: Understand your legal responsibilities

Now that we have put some flesh on the idea of a data situation it should be apparent that the movement of data across multiple environments complicates the question of who is responsible for data and more specifically: what is your role in respect of those data? Are you data controller, processor or user? The key to resolving this is to: (i) *know where the data have come from* and under what conditions and (ii) *know where they are going* and under what conditions. The 'conditions' you need to take account of are:

1. The status of the data in each data environment in the data situation, whether they are personal, de-identified or anonymous data.
2. The data provenance, i.e. who decided to collect the data (including what data and who it is about), established the legal grounds for doing so and determined the means for processing it.
3. The enabling conditions for the share or release of the data (in an anonymised form), i.e. how is that processing fair and lawful?
4. The mechanism for a data share or release, e.g. a data sharing agreement or contract, or an end user or open licence.

Despite the complexity of the questions here, many situations can be subsumed under two common models of processing responsibilities, which we will outline shortly and which cover questions 1-4.

Data status: are my data personal?

To consider this, let us remind ourselves of the definition of personal data given in chapter one. Personal data are 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.

Article 29 Working Party's 2007 opinion on 'personal data' identifies four main building blocks underpinning this description. It is worth looking briefly at these building blocks as they make explicit some of the inherent ambiguities that stem from applying an abstract concept to real world situations. They are as follows: 'any information'; 'relating to'; 'identified and identifiable' and 'natural persons'. We shall look at each in turn.

Any information – The Working Party notes (and we concur) that this phrasing calls for a wide interpretation, and includes any sort of statement about an individual. The statement may be objective, such as someone's health, employment status or

qualifications, or subjective, such as an opinion or assessment like 'John Smith is a good employee'. For information to be personal data it need not be true or proven.

Relating to – This means the information is about an individual. The relation may be direct, for example their exam transcript in their school file, medical test results in their hospital record or a CV in their employment file. This is clear, but when the relation is indirect it can become complicated because, depending on the circumstances, indirect information may or may not be personal data. For instance, the valuation of a property (directly about an object, of course, not a person) may be considered as personal data if it is used to determine the way in which that person is treated, if for example it is used for setting the level of taxes. 'Data relates to an individual if it refers to the identity, characteristics or behaviour of an individual or if such information is used to determine or influence the way in which that person is treated or evaluated' (Article 29 Data Protection Working Party 2007:10).

Identified and identifiable – A person is considered to be identified if within a group of persons they can be distinguished from all others in the group. A person is identifiable where the conditions exist to identify them. As discussed at some length in chapter 2, within the ADF we consider whether a person is identifiable or not to be heavily contextualised. The advantages of this approach is that it disambiguates the technical processes that you will need in component 7; the cost is that you will need to do more work at this stage identifying who the agents are and whether the data are personal for them.

Lastly, the concept of **natural persons** – The protection afforded by the rules of the Directive applies to natural persons (that is, to human beings) and more specifically living persons. In some circumstances, there are two further legal considerations that extend this in the UK. The first concerns the Statistics and Registration Services Act (2007), which expands the protection to include any body corporate for the purposes of official statistics. The second concerns medical records of deceased persons. As we know the DPA protects the personal data of living persons only as deceased persons under the Act are no longer considered data subjects. Although there are no clear legal obligations of confidentiality to deceased persons in the UK, for medical data the Department of Health and the General Medical Council have deemed that there is an ethical obligation to ensure that confidentiality continues to apply to these data after death. This is supported by the Scottish Freedom of Information Act (2002, section 38) which classes medical records of deceased persons as personal data.

To summarise, there are no unequivocal rules about how to determine what constitutes personal data. However, by working through the four key components of the definition in the DPA and the EDPD ('any information', 'relating to', 'identified and identifiable' and 'natural persons'), it is usually fairly straightforward to make a decision one way or the other. We shall now move on to consider in more detail the issue of processing roles.

Model 1: Single controller

This is the simplest model of data processing responsibilities.

Imagine that Barsetshire LA decides to collect and hold personal data from its service users, and determines the legal basis for this. It agrees to share a subset of the personal data with the Agency of Public Sanitation (APS) for the purpose of supporting APS's public service work. Barsetshire has procured its service users' consent for the share. The share is formalised with a written contract[73] which stipulates how APS can process the data, specifying how the data can be used, whether they can be (further) disclosed, under what conditions and to whom. Under this contract, APS's processing responsibilities are limited to determining how it will hold the data and keep it secure.

The model described here, in Figure 3.4, illustrates the most straightforward processing relationship between two organisations. Barsetshire is the data controller, having determined the manner in which, and the purposes for which, the data are processed. As such it retains overall responsibility for the data. APS's responsibilities are related to ensuring it does not breach the conditions of its contract with Barsetshire and particularly its data security undertakings. However the legal responsibility for compliance with the DPA falls directly on Barsetshire (the data controller) not APS (the data processor). Barsetshire cannot pass on its responsibility to APS and has a duty to ensure that APS's security arrangements are at least equivalent to its own as well as taking reasonable steps to ensure that these are maintained, for example by regularly auditing APS. In terms of enforcement, even if APS were considered negligent because, for example, it did not follow agreed security measures, the ICO cannot take action against it, although Barsetshire could pursue a civil action for breach of contract. On the other hand, if APS were to

[73] The DPA requires that when a controller discloses personal data to a data processor it uses a written contract rather than a data sharing agreement. This is so only the controller can exercise control over the purpose for which and the manner for which personal data can be processed.

deliberately use the data for its own purposes (which would also break the terms of its contract with Barsetshire), it would both become a data controller in its own right and is likely to be in breach of the first principle of the DPA and the ICO could take enforcement action against it (see UK: Information Commissioner's Office (2014a) *Data Controller and Processor Guidance*).

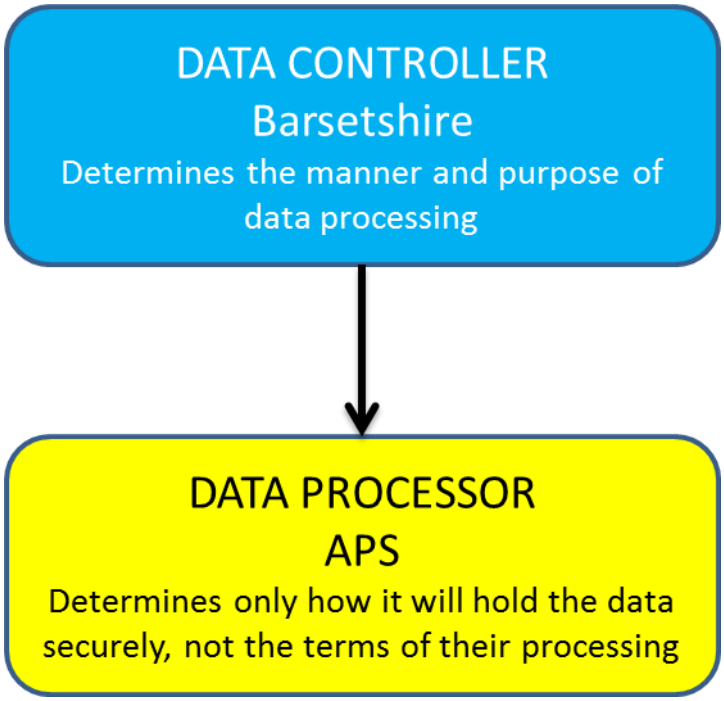


Figure 3.4: Model 1

Model 2: Pluralistic control ('jointly with others')

Our second is a more complex model of data processing responsibilities and involves situations where there is more than a single data controller.

Pluralistic control involves data controllers working together in different forms and combinations (Article 29 Data Protection Working Party, 2010). The notion of *together* may involve data controllers acting jointly and equally in determining the purpose and means for a single processing operation or it may be more complex than this; it may take the form of a looser relationship where data controllers share only purpose or means. Let us look further at this using an example.

Imagine that the Department of Social Affairs (DoSA) along with the Local Authorities (LA) of Barsetshire, Rabsetshire, Arbsetshire (and all other LAs in England and Wales) determine the purpose and legal basis for collecting data on public transport usage at an individual level. The DoSA determines the means, including determining what data should be collected and the manner in which it

should be stored and managed (via a shared portal), as well as who can have access to it. Because the LAs are involved in determining the purpose of the processing operation (although not the means) they are considered joint data controllers with the DoSA, for the data in question. As the sharing is systematic and large-scale, all parties sign up to a data sharing agreement covering the means of the processing operations. This agreement[74] stipulates that the DoSA takes responsibility for establishing and managing the shared portal whilst the LAs take responsibility for collecting and uploading their data to it. This ensures that it is clear who is responsible for compliance with the DPA for which (and all) aspects of the processing operation.

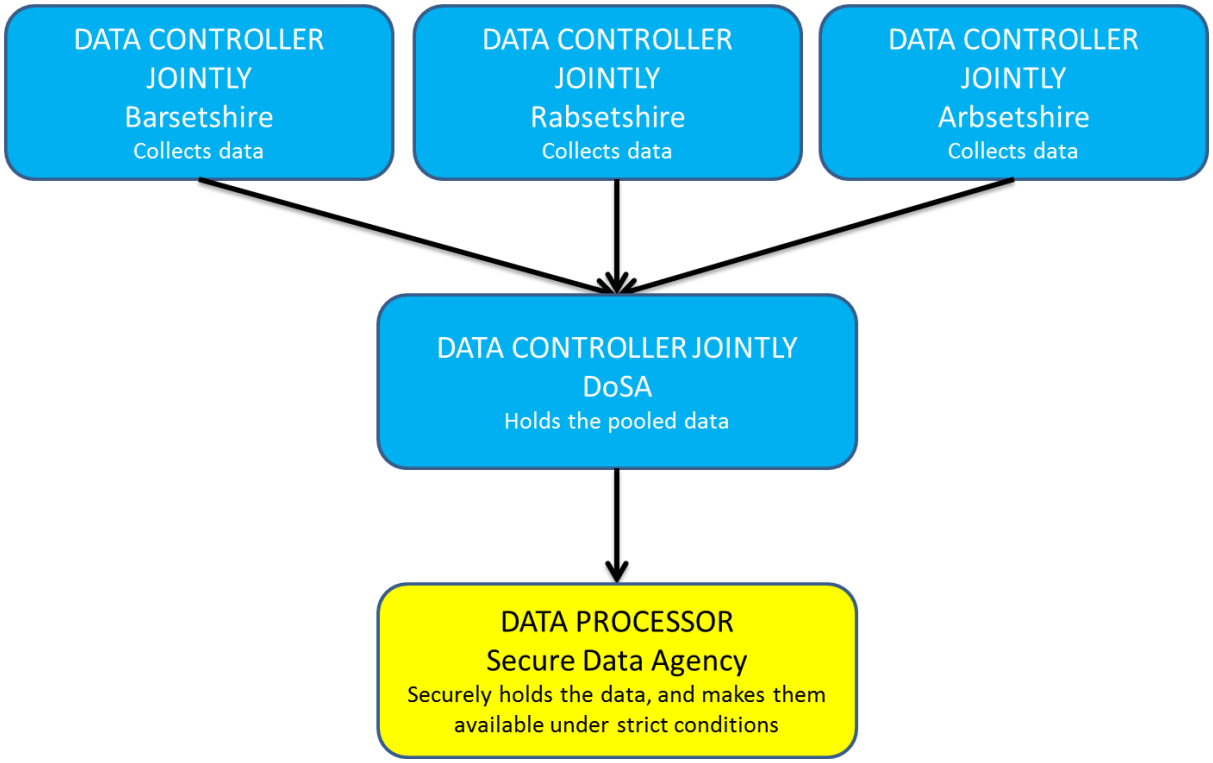


Figure 3.5: Model 2

Let us extend this example further and imagine that, as outlined in the data sharing agreement, the DoSA makes a de-identified subset of the data available to approved accredited researchers (because there is a compelling public benefit case for reusing the data for research purposes). The data are made available in a secure lab run by a third party. Under the control conditions of the secure lab, the data these researchers will access are functionally anonymised. The lab, like APS in the previous example,

[74] Parties acting jointly have a degree of flexibility in distributing and allocating responsibilities among themselves as long as they ensure full compliance. The distribution of responsibilities needs also to be reasonable in order to be enforceable. For further information, see EU: Article 29 Data Protection Working Party (2010) and UK: Information commissioner's office (2014a).

has processing responsibilities for the de-identified data which are formalised in a contract stipulating the conditions under which they should be securely held and accessed by researchers. The DoSA and LAs retain full data protection responsibilities.

Of course there are many variations of the pluralistic model. The key is to remember that even where data processing relationships are complex, responsibilities for compliance with the DPA should be clearly allocated. In determining how best to do this the general principles remain: *know where the data has come from* and *how and why it was collected and* under what conditions and *know where the data is going* and under what conditions.

Component 3: Know your data

When thinking about whether, and how, to share or release your data safely, a key consideration will obviously be the data themselves. In this section, we set out a top level examination of data focusing on the data's type and properties. Identifying these features will be relevant for both components 6 and 7 of the ADF later. Also it is possible at this stage to make some straightforward decisions which will simplify the more detailed processes you will go through later. But the main purpose of this component is to get a picture of the data in much the same way that a data analyst might explore a dataset before starting to build a multivariate model.

As we will illustrate, the data features *status*, *type* and *properties* are central to the issue of anonymisation, whilst other points are useful indicators as to the general level of risk[75] you might assign to your dataset. In more detail:

1. **Data subjects**: Who are the data about and what is their relationship with the data?
2. **Data type**: What form are your data in, e.g. statistics or text? What level of information about the population do the data provide, e.g. are they microdata or aggregated?
3. **Variable types:** What is the nature of each variable within the data? For instance, are they direct identifiers, indirect identifiers or targets?
4. **Dataset properties**: What are the top level properties of the dataset, e.g. its age, quality, etc.?

[75] We use the term 'general level of risk' to describe a system for categorising risk. It provides no more than a guide about risk levels i.e. low risk, medium risk, high risk. This is considered further on in this section.

We consider now each of these features in turn and in doing so highlight their relevance to the question of anonymisation.

Data subjects

In most cases, who the data are about is a straightforward question. However, as we demonstrated in section 2.1, data can be indirectly about people when they are directly about something else (in the example we used the topic was house fires). However, you should also be mindful that data which is directly about one group of data subjects may also be indirectly about another group; for example patient record data for a particular GP practice could indirectly be about the practice's GPs (e.g. their prescribing practices).

You should also consider here whether the data subjects in the collective represent a vulnerable group and the extent to which the data subjects have given consent to any of the data processing involved in your data situation and/or the extent to which they are aware of it. We will consider these issues more centrally in component 5.

Data type: what type of data do I have and what type of data should I share/release?

If you have collected your own data about people then it is likely to be in the form of individual unit records, or *microdata*. Such data are commonly stored in digital databases as single records of information where the rows represent a single population unit (person, household, etc.) and the columns represent the information (variables) you have collected about them. For the purpose of sharing or releasing data you may decide not to make available an anonymised version of the microdata, but instead to aggregate your data and make it available as an anonymised table, graph or map. To assist you in making decisions about what type of data to share or release let us consider the particular disclosure risks associated with each.

For aggregate data, particularly for small geographical areas, such as for example a census output area or postcode sectors, attribution disclosure[76] and disclosure by differencing are considered to be particular problems. See Smith and Elliot (2008) and Duncan et al (2011) for a discussion of these problems.

For microdata, re-identification disclosure is a particularly challenging problem. This is because of the difficulty of determining which variables or combination of

[76] This is where there are zeroes present or where they are inferable in some combination of the variables in the aggregated data.

variables might make an individual unique in a dataset and therefore stand out as vulnerable to re-identification (we consider this further in the next subsection).

Variable type: what types of variables are in my dataset?

Variables can be direct or indirect identifiers, or targets. Most datasets will have a mix of all three types.

A target variable is usually one that a reasonable person would consider to be sensitive and that is not widely available. Identifying the target variables will inform you about likely harm that will arise from a disclosure and may also inform the construction of disclosure scenarios (see component 6).

At this stage you are not attempting to form fully specified scenarios but simply explore the data, sorting variables into appropriate types. Some variables might be obvious identifiers – for example sex and age are routinely included in most key variable sets; others you may not yet be sure about. The purpose of this is to get some idea of the scope of the anonymisation problem.

As described in section 2.2.1, direct identifiers are any attributes or combination of attributes that are structurally unique for all persons in your data, such as unique reference numbers like NHS numbers or social security numbers. In most dynamic data situations you will suppress (or possibly pseudonymise) the direct identifiers as a first step. Therefore, being clear about which variables are direct identifiers is important.

An indirect identifier in contrast can be any attribute (or set of attributes) that, whilst not structurally unique, are likely to be unique for at least some individuals in your dataset and in the world. An example of indirect identifiers might be the combination of age, marital status and location variables. Whilst these are not immediately obvious identifiers, if we return to the example of the sixteen year old widower and imagine one is living in rural Scotland this rare combination of attributes is likely to make him unique and thus at greater risk of re-identification. The important point is that rare combinations can crop up and create a risk of someone spontaneously re-identifying them.

Sensitive data – Sensive data is thought to increase re-identification risk because (i) it is more likely to be targeted because it is interesting, and (ii) the impact (and potential harm) of a disclosure may be greater.

'Sensitive data' encompasses a wide range of information and holds a special position in the DPA which means that even when data are being processed for secondary use, if they are classed as sensitive an additional reason to process the data is required (see Schedule 3). Within the DPA, the definition of 'sensitive' is based on a list of categories which are widely regarded to be an incomplete list of what might intuitively be covered (financial data for example is notably absent).[77] So you may consider that there are other categories of data which are sensitive and identifying. For example, McCullagh's (2007) research on the issue of data sensitivity suggests that the current categories used by the DPA need to be brought up to date with 21st century developments in areas such as IT and biometrics. However, rather than adding more categories to the list of what is sensitive data, McCullagh argues for a different approach focusing on harms or likely harms.[78] Relatedly, Simitis contends that personal data becomes sensitive according to its context, arguing that 'any personal datum can, depending on the purpose or the circumstances of the processing be sensitive'(1999:5) So one's address is for most people a fairly innocuous piece of information but for somebody on a witness protection scheme it becomes highly sensitive. This line of reasoning implies that you should think carefully about all categories of data you plan to share or release.

The overarching point is that if you are dealing with sensitive data then the risk is higher both in terms of the likelihood of a deliberate attempt to access the data and the impact of an attempt if successful. As well as impacts on the data subjects, unintended disclosure of sensitive personal data is also likely to be more damaging to the data controller than disclosure of non-sensitive data, because the impact on public trust and reputation is likely to be greater. If you do not have the data subjects' consent to release or share an anonymised version of the data then the risk to your organisation's reputation if there was a subsequent breach is amplified further. To put it another way, if you do not have consent then your overall data situation is more sensitive.

[77] In the EU member states, the list will be expanded when the European General Data Protection Regulation comes into force.

[78] It is noteworthy that this notion has begun to gain traction in some jurisdictions. For example, the recent Singaporean personal data protection act (2012) includes a provision that 'An organisation may disclose personal data about an individual… if the disclosure is for archival or historical purposes and if a reasonable person would not consider the personal data to be too sensitive to the individual to be disclosed at the proposed time'. (Schedule 4 clause 1 r).

Dataset properties: what are the properties of my dataset?

The properties of a dataset can *potentially* increase or decrease the risk of disclosure. We say potentially because, as acknowledged, we are at this stage doing no more than getting to know data without reference to the data environment.

The use of a general risk indicator, such as the one described below, merely acts as a guide to help you think about which data properties require particular attention when you are doing further analysis. It is not a substitute to the requirement for a careful analysis of your data, but a precursor.

Data properties include: quality, age, data detail, coverage and structure. We look now at each of these and outline how and why they may affect disclosure risk.

Data quality: It is generally accepted that all data will contain some level of error. Error can originate from the data subject, data collector and/or the data collection process (Bateson 1984). The aim as a data holder is to ensure that the error level in your data is small; after all there is little point in sharing or releasing data that does not represent whatever it is supposed to represent because it is distorted by the (low) quality of the data.

However, ironically a small level of error, inherent in all data, has some advantages as it offers some degree of natural data protection (Duncan et al 2011).

Age of data: It is generally understood that the older the data[79] the harder it is to identify people correctly from them. This is because people's information may change over time as, for example, they move location, change jobs or get married. Thus older data may acquire a basic level of data protection because of the issues associated with divergence as discussed in chapter 1.

Hierarchical data: This is data that contains information for members of a group who are linked with one another and is a common source of disclosure risk in business data. The data are considered more risky because they provide (more) information that might make a data subject unique in a dataset and as such potentially identifiable. For example, the combination of age and sex of all members of a household will be unique for most households above a relatively modest size (Duncan et al 2011).

[79] The other (reverse) problem with older data relates to an increased risk of associating incorrect information with people identified within a dataset because the information is out of date.

Longitudinal data: This is data about a defined population which is collected over time and linked. These are considered riskier because of the potential to capture potentially unique changes in information over time such as changes to one's marital, economic, employment and health status and location that stand out amongst other longitudinal patterns. Again this may increase the likelihood of a data subject being unique in a dataset and as such potentially identifiable.[80]

Population or sample data: Population data includes census data and data for all people in a particular group such as benefit claimants or hospital patients. It is considered more risky because there will be little uncertainty as to who is represented in the dataset.

Capturing the data features

In Appendix E you will find a template for capturing the above features and perhaps recording any top level actions that you might make. For example, you could decide that you will release a flat rather than hierarchical file or that date of birth will be recoded to single year of age. These decisions simplify the technical work required (in components 6 and 7). Your framing for this capture of features will be the use case, to which we now turn.

Component 4: Understand the use case

In determining the use case for your data you need to understand three things:

1. **Why:** Clarify the reason for wishing to share or release your data
2. **Who:** Identify those groups who will access your data
3. **How:** Establish how those accessing your data might want to use it

Working through these three points will help you with decisions about both *what data you can safely share or open* and *what is the most appropriate means by which to do this*.

Firstly, you should be clear about your reason(s) for sharing/opening, because your actions will:

1. Require resourcing which, in all likelihood, you will need to justify.

[80] Analytically, longitudinal data could be treated as longwave – i.e. the slowest form of – dynamic data (data that updates over time). In practice however they are analysed as static datasets and therefore in most data situations the longitudinal element is treated as another property of the dataset. It does however lead to some different intrusion scenarios as it is necessary to consider the likelihood of an intruder also having access to longitudinal data.

2. Carry a risk so you need to be able to perform a rigorous cost/benefit analysis.

There are numerous reasons for disseminating data. Perhaps it provides useful information for stakeholders or about your organisation, offers new insights/perspectives on a topic, offers a benefit to particular groups, supports the more effective/efficient use of a service, or maybe you have received an FOI (freedom of information) request. Thinking through why you are disseminating your data automatically brings in the other two questions, the 'who' and the 'how' of access.

Your potential users may be a single organisation, a defined group or several different user groups. You may decide to provide different data products via different dissemination routes.[81]

Direct consultation with your potential data users is one method for understanding the use case and can take many forms. Whilst it is not within the remit of this chapter to talk about user engagement in detail it is worth noting that a variety of methods is available such as interviews, focus groups, web surveys or a call for written feedback, the last of which you could administer directly through a website or via a third party. The exact nature of the type of activity you might carry out will depend on the number and type of users and the drivers of the programme to share/release. Are they internal or external to your organisation? Are you responding to a contractual or statutory obligation, or are you trying to increase the utility of your data? Is it a drive for transparency and good will, or do you hope to provide an income stream?

However you decide to engage with your users, it is helpful from the outset to identify who they are and how they will use your data, although this is not always possible as the use case may emerge over time. Certainly data released for one reason and for a particular user group may over time be used serendipitously for purposes not first envisaged and by new groups of users. Whilst you may not be able to initially determine all possible uses for your data, you should try to keep abreast of how they are being used. How you can go about doing this is discussed in component 9 below.

[81] If you make available different data products via different dissemination routes you will need to take account of the risk of disclosure for each in combination with the others. See component 6 for further discussion.

That there will be some benefit to the reuse of data is axiomatic in today's 'big data' climate. The demand for data seems insatiable. So clarifying the questions to be answered by your data, or what needs it is hoped they will meet, is a good place to start when thinking about exactly what data to release and how it should be specified.

Once you have determined the sort of data product that your users want or need (or what data product is likely to be useful to a wider audience), you then have to think about how best to share or release it. Remember the central objective is to disseminate safe and useful data. There is a trade-off between risk associated with the environment and the utility of the data themselves. Broadly speaking the less controlled the access environment, the less detailed and less useful (all things being equal) the data must be in order to ensure safety. Let us consider briefly some of the options for sharing and releasing data.

Data sharing: This may (though not always) involve the movement of data from one organisation to a partner or associated organisation where there is some sort of established relationship between those organisations. Whether or not there is already a relationship between the organisations proposing a share, data shares should always be formalised and managed using a contract or sharing agreement which (i) makes clear who is responsible for what and (ii) ensures fair processing, usage and retention of the data.[82] By using a contract or DSA you can manage (some of) the disclosure risk associated with the data share.

Data release options: See chapter 2 for discussion of this topic. Suffice it to say that whilst there are several data release options available, which one you choose depends on the data you plan to release, their sensitivity, and the proposed usage of them. As a general rule, the more you restrict access to your data the greater control you have over how they are used. Conversely, a more liberal regime usually means you relinquish more control, and hence you need to think about restricting the (detail in) data themselves. Allowing greater access to your data does not automatically produce high risk, as long as your data are sufficiently anonymised given the release environment.

It is worth noting that in applying the risk-utility concept[83] you will need to think beyond the impact of anonymisation techniques on your data. Think also about what

[82] See UK: Information Commissioner's Office (2011) for the UK regulator's position on this.
[83] See Duncan et al (2001) for the original exposition of the risk utility framework.

the application of a particular technique might mean to your users. For example, complex methods may not be appropriate for data that you plan to make widely available because non-specialist users may not understand their impact on the data. We discuss this issue below in component 7.

To recap, establishing the use case for your data will help you think about what data you could share or release and how to do that safely. In determining these things you will need to balance data utility and data protection. This balance is a well-recognised trade-off and what essentially makes the task of producing safe, useful data challenging.

Component 5: Meet your ethical obligations

As we stated in chapter 2, acquaintance with the ethical issues related to the reuse of data should not deter you from sharing or releasing data. We outline in this component how you can go about meeting your ethical obligations whilst maximising the value of your anonymised data.

Consent and other means

Where possible, seek consent from data subjects for what you intend to do with their data once anonymised. Although, as we stressed in chapter 2, consent is not a panacea you are in a much stronger position ethically if you have it than if you do not. However, if seeking consent, do think carefully about the assurances you give to data subjects about what will happen to their data. You should respect the assurance you give because, if your organisation gives certain assurances to your data subjects and then breaks them, then you are not processing data fairly and therefore are in breach of principle 1 of the DPA.

When consent has not been sought, you might want to consider being as transparent as possible, and engaging with stakeholders where practicable.

Transparency of actions

To be transparent, at the very least explain simply and clearly to your data subjects how you reuse data with a description of your rationale. This could be done for example on your website, during any public facing event, or in relevant publications, or you could undertake to explain on request to interested parties.

Stakeholder engagement

Consulting with your stakeholders is a useful exercise, an effective way of understanding your data subjects' views on your proposed data sharing/release

activities and addressing their concerns. However, it can be resource intensive and so you might consider undertaking it only after you have run through other options and if there is potential for concerns to arise. It is also worthwhile looking at how similar organisations in your sector are sharing data and whether any concerns have been raised about their practices. Finally a growing amount of survey and focus group work has been done on data subjects' views on data sharing and reuse, particularly in the health sector; we recommend that you look at this to help inform your thinking.

The importance of good governance

Key to ensuring you meet your ethical responsibilities either as a data controller or data processor is good governance. On a broad level governance is about the organisation of your data processing activities formalised in principles, policies, and procedures for data security, handling, management and storage, and share/release. To underpin this you should have a clear picture of what the flow of data looks like within your organisation and what your processing responsibilities are.

In practical terms this includes (but is not limited to) the following factors.

Governance and human resources

- Identify a person in your organisation who will be responsible for authorising and overseeing the anonymisation process and ensure that they have the necessary skills and knowledge to do this.
- Ensure that all relevant staff are suitably trained and understand their responsibilities for data handling, management, sharing and releasing.

Governance and internal structures

- Establish principles, policies and procedures for acting as a data controller.
- Establish principles, policies and procedures for sharing data including how you will monitor future risk implications for each share (see component 10).
- Establish principles, policies and procedures for releasing data including how you will monitor the future risk implications for each release (see component 10).
- Establish a comprehensive record-keeping system across all your operational activities related to your data protection policies and procedures to ensure there is a clear audit trail.
- Undertake a Privacy Impact Assessment (PIA) for all your data products and/or across your organisation as a whole.

- Establish principles, policies and procedures for identifying and dealing with cases where anonymisation may be problematic to achieve. You should also consider at what point in the process (in dealing with a difficult case) you should seek external help and advice from bodies such as the ICO or expert groups such as UKAN.
- Establish principles, policies and procedures for dealing with data breaches. Depending on your organisation's particular needs you may choose to develop separate policies related to different potential data breaches, or develop a single policy. Whichever you chose you will need to consider how a breach might occur and how you will respond to it.

The what and how of a data breach

1. Define a data breach.
2. Identify the types of data breach relevant to your data situation.
3. Identify those factors likely to lead to a breach, such as the loss of an unencrypted disc taken out of the workplace or the accidental emailing of data to the wrong person. Thinking through a range of possible breach scenarios can be very useful in helping you identify how a breach might arise from your usual processing activities, as well as what errors, procedural violations or malicious intent may also occur.
4. Establish measures to limit/avert those factors likely to lead to/facilitate a breach.
5. Establish how you will address violations of these measures.

Responding to a data breach

We address this issue in detail in component 10 of the framework below, but note that it includes the following areas:

1. The containment of a breach.
2. Assessing and dealing with any ongoing risk.
3. Notification of a breach.
4. Review and learning lessons.

Governance and horizon scanning

- Keeping up-to-date with any new guidance or case law that clarifies the legal framework surrounding anonymisation. For example you could regularly

view the UKAN website[84] which provides information on anonymisation, and the ICO[85] website which provides information on data protection.

- Talking to other organisations in your sector to share best practice. You might want to consider going to events such as the ICO Annual Data Protection conference to keep up to date with current issues and networking with other people working in data protection.

Ensure Privacy Impact Assessment is embedded in your organisation

It is considered best practice to think about and embed privacy into the design of your data processing activities right from the start. Although it is not a legal requirement under the DPA to undertake a Privacy Impact Assessment there are many good reasons for having one when you process data. It will (i) help you be aware of and address any particular privacy issues, (ii) ensure the transparency of your activities, (iii) promote trust in what you do, and (iv) help you to comply with the DPA and any other relevant legislation. For further information see the ICO's Guide to Conducting Privacy Impact Assessments code of practice (UK: Information Commissioner's Office; 2014b).

3.2 Disclosure risk assessment and control

Risk assessment and control should usually be an iterative, not linear, process. There is rarely a single possible solution; the risk analysis might suggest changes to the data specification which, once experimentally applied to the data, require a fresh risk analysis. Furthermore, there are several types of risk assessment, and you should be strategic in how you apply them. Some are quite resource intensive and therefore should only be applied to near-final versions of the data if they are needed at all (assuming your budget is limited).

This process will be constrained by the use case and the resources available. As ever, our goal is to produce data that meets the requirements of the use case. The use of resources to address potential problems should be proportionate to the impact of a breach.

[84] www.ukanon.net
[85] www.ico.org.uk

Component 6: Identify the processes you will need to go through to assess disclosure risk

Risk assessment is a crucial step in the process of producing safe useful data, helping you to:

- determine whether your data should be shared or released at all;
- determine how much disclosure control should be applied; and
- think about the optimum means for sharing or releasing your data.

In practice this can be very complex and risk assessment is probably the most difficult stage of the anonymisation process, requiring judgement and expertise on the part of the data practitioner. The complexity is partly because it is not evident what additional relevant information might be taken into account and how different factors might affect risk. As such factors include the motivation of the intruders, the efforts to which they might go, and the techniques they might use, it is clear that such factors can never be definitively specified. It is also unknowable what information might become publicly or privately available in the future, from sources other than yourself, which might be used to link with the data you wish to release to reveal identity. Notwithstanding these inherent limitations, there are steps you can take to assess the disclosure risk associated with your data and the share/release environment.

We introduce a four-part process for assessing disclosure risk. The first two procedures are always necessary, while the third and fourth may or may not be required depending on the conclusions drawn after conducting the first two.

1. ***Incorporation of your top level assessment to produce an initial specification.***
2. ***An analysis to establish relevant plausible scenarios for your data situation.*** When you undertake a scenario analysis, you are essentially considering the *how, who* and *why* of a potential breach.
3. ***Data analytical approaches***. You will use data analytical methods to estimate risk given the scenarios that you have developed under procedure 2.
4. ***Penetration testing,*** which involves validating assumptions made in 2 by simulating attacks using 'friendly' intruders. The ICO recommends carrying out a *motivated intruder test* as part of a practical assessment of a dataset's risk. This can be both informative and good practice but takes skill and expertise as well as time and resources.

Incorporating your top level assessment

We described in component 2 how to undertake a top level assessment of disclosure risk by identifying those features of your data that can potentially increase or mitigate risk. Let us remind ourselves of those features.

Data quality	May offer some data protection
Age of data	Older data are less risky
Hierarchical data	Increases risk
Longitudinal data	Increases risk
Population data	Increases risk. Conversely sample data offers some protection.
Sensitive data	Potentially increases the risk and impact of a disclosure
Key variables	The core of the re-identification problem
Microdata	Re-identification disclosure is a particular problem
Aggregate data	Attribution disclosure and disclosure by differencing are particular problems

Table 3.1: Risk-relevant features

This top level analysis enables you to identify where you need focus your attention in the technical analysis that follows. At this stage you can also simplify the dataset. Anything that you can do to reduce the complexity of the data will in turn reduce the complexity of the technical analysis that you have to conduct at the next stage.

It might be now that you identify a sensitive variable that is not needed for the use case – take it out. Your default assumption should be that if it is not needed then it should be deleted. If the data are hierarchical, is the preservation of that property required for the use case? Being hierarchical will often magnify the risk markedly, and you may have to compensate with some heavy controls elsewhere in the data. Could the data be simplified to a non-hierarchical structure?

Are there any variables with a lot of detail? If so, is that much detail really necessary for the use case? Frequency tables and descriptive statistics also need to be looked at. Are there variables whose distribution is highly skewed – say with one category which contains most of the cases and a dozen small categories. Can the small categories be merged? Are there any continuous variables on the dataset which might be rounded or banded?

Such brutality to the data may seem blunt or even draconian but remember that whatever you do you will not eliminate the risk. The more data you release, the

riskier it will be, so if a risk is unnecessary for the use case, do not take it.[86] Initially, removing low-utility/high-risk features will not impact on the overall utility. Eventually though you will hit a point of diminishing returns, where a utility reduction will start to become evident and then it is necessary to move on to the second procedure.

Scenario analysis

As outlined in chapter 1, the purpose of scenario analysis is to ground your assessment of risk in a framework of plausible events. If you use the Elliot and Dale framework outlined in section 2.3.1 then you will run through a series of considerations using simple logic to arrive at a set of key variables. In constructing these you need to consider all the sources of the data that the would-be intruder might have access to. Below are examples of other sources of data that may be relevant when developing your scenarios of disclosure.

- ***Public sources of data:*** including public registers, professional registers, electoral registers, land registry, estate agents' lists, newspaper reports, archived reports and announcements, parish records and vital statistics such as birth, death and marriage records.
- ***Social media and other forms of found data:*** including data generated by the data subjects themselves in online interaction and transaction ('found data'). This runs from deliberate self-publication (CVs, personal websites), to material where the goal is primarily interactive (social networking sites). Needless to say this is a growing source of publicly available information and Elliot et al (2015) demonstrate that it is plausible to attack an open dataset using a combination of social media data and other publicly available sources.
- ***Other similar data releases:*** including releases from your partner organisations and other organisations in your industry or sector.[87]

---

[86] One factor to bear in mind here is the nature of sharing/dissemination you are working with. If you are in a data situation where you will be dealing with a series of multiple bespoke data requests, then performing a risk analysis and editing process with each individual request could be quite onerous. It may be simpler to produce a single dataset that meets the negligible risk criterion in any conceivable use case. However, the disadvantage of this approach is that it will inevitably be a lowest common denominator dataset and some users may not be able to access the data they want. As ever there is a balance to be struck here.

[87] Ideally if multiple organisations were releasing open data on the same population then they would co-ordinate their anonymisation processes. However, in most cases in practice, such an undertaking will be very difficult. The importance of other releases will be greater if the data generation processes are similar, the time of collection is similar and if there is partial overlap of variables. This set of

- ***Official data releases:*** including data releases from the Office for National Statistics (in the UK), government departments and agencies and local authorities.
- ***Restricted access data sources:*** including the resources of any organisation collecting data. At first it may seem difficult to imagine how you would know what is in such data sources but they can often be the easiest to find out about. Why? Because although the data are hidden, the data collection instruments are often public. They include the forms that people have to complete in order to access a service, join an organisation or buy a product. If you can access the forms used to gather data, then you can make a pretty good guess about what data are sitting on the database that is fed by the form. For the task of generating key variables that should be sufficient.

It is easy to become overwhelmed by the feeling that there is too much data out there – where do I even start? Certainly doing a full scenario analysis is very time-consuming and beyond the resources that many organisations are likely to have available. Fortunately, for many data situations a full analysis will be disproportionate. This will be particularly true where you are working in a tried and tested area. If other organisations have been releasing similar data for a while without any apparent problems then your resources that you need to devote to this element can be more modest.

One tool that can cut down the amount of time required at this stage is the standard key set. Standard keys are generated by organisations carrying out ongoing data environment analysis (scanning the data environment for new data sources). You should be aware that standard keys are generic and are set up primarily for use with licence-based dissemination of official statistics and will not be relevant to every data situation. However, the standard keys can be useful because, if your data are not safe relative to these standards, then in itself that indicates that you may have a problem, even before you consider non-standard keys. A set of standard keys can be found in Appendix A.

So what in practice can you do if you are not carrying out full scale data environment and scenario analyses? The simplest approach is to carry out thought experiments that make the imagined adversary more specific.

---

circumstances will usually only arise when the organisations are closely related. This, in principle, allows *key extension*; as both datasets are anonymised we are not here talking about direct re-identification, but the fusion of two anonymised datasets could make both more vulnerable.

For example, imagine that you are a local authority wanting to release a dataset of social care service users as open data. Suppose the dataset contains 7 variables: age (banded), sex, ethnic group, ward (LA subdivision), service accessed, the year that service was first received, and type of housing.

Now imagine a data intruder who draws on publically available information to attack a dataset that you have released as open data. Run this scenario through the Elliot and Dale framework. In particular, think of a plausible motivation and check that this passes the 'goal not achievable more easily by other means' test. In this example you might end up with inputs something like this:

- ***Motivation: what is the intruder trying to achieve?*** The intruder is a disgruntled former employee who aims to discredit us and, in particular, our attempts to release open data.
- ***Means: what resources (including other data) and skills do they have?*** Publically available data, imagine that they are unemployed, have unlimited time, but do not have access to sophisticated software or expertise for matching.
- ***Opportunity: how do they access the data?*** It is open data so no problems at all.
- ***Target variables:*** which service(s) individuals are using.
- ***Goals achievable by other means? Is there a better way for the intruders to get what they want than attacking your dataset?*** Possibly, but discrediting our open data policy would be effective.
- ***Effect of data divergence.*** We believe our dataset to be reasonably accurate. However, we are only publishing data that is at least one-year-old. The intruder's data will be less reliable. This will create uncertainty for the intruder but not enough to rely on.

Once you have a plausible scenario then look through the standard keys set to see if any of those correspond meaningfully to the total information set that the intruder might have. In this case the standard key B4.2 looks relevant. If we cross reference the list of variables under that key with the list that we are considering releasing, that gives us the following intermediate outputs:

- ***Attack type: what is the technical aspect of statistical/computational method used to attack the data?*** Linkage of data about individuals living within our local authority derived from publicly available information to records in the open data set.
- ***Key variables:***

- Ward
- Ethnic group
- Age (banded)
- Sex

These key variables can then be used as a starting point for the technical disclosure risk assessment. If you are taking this approach then it is wise to construct more than one scenario. The number that you will need will depend on the totality of the data situation and specifically who will have access to the data and the complexity of the data in question. With this situation we are talking about open access but relatively simple data. With open data we will often want to also assess the nosy neighbour scenario (see Elliot et al 2016 for a rationale for this), which would suggest adding *type of housing* to the list of keys, but would also mean that we were simulating an attack by an unsophisticated intruder who was just trying to find a single specific individual (rather than any high-certainty match from a host of possibilities).

Of course if your data does not nicely fit into the format of the standard keys then you are going to have to do some work to populate this framework yourselves. You should avoid focusing too closely on apparent vulnerabilities in the data. For a good analysis of the pitfalls of doing this, see, for example, Sánchez et al's (2016) critique of de Montjoye et al's (2015) account of the uniqueness (called 'unicity' by de Montjoye et al) of small strings of credit card purchases. Uniqueness – and particularly data uniqueness – does not in itself re-identify anybody. Uniqueness does indicate vulnerability but if there is no well-formed scenario through which that uniqueness can be exploited then no re-identification can happen. On the other hand a sophisticated intruder might focus on those vulnerabilities to carry out a fishing attack. It comes down to whether there is a well formed and feasible scenario where they would be motivated to do that.

So create your scenarios, generate your key variables and then carry them through to your risk assessment.

Data analytical risk assessment (DARA)

Having gathered the low hanging fruit of data reduction, and generated your sets of keys now you are ready to move on to carry out a *data analytical risk assessment* (DARA). We would always recommend that you get expert advice at this stage even if only to ratify what you have done. However, much can be done without external help, and the more that is done in-house, the richer the conversation that you can have with independent experts, including communicating your specification of the

problem to them, and interpretating their findings and recommendations. In this section, we will set out a process that could be performed in-house, without (i.e. before) consulting anonymisation experts.

File level risk metrics

The first step in the DARA is to obtain a file-level measure of the risk. There are quite a few of these and selecting the right one can be a bit of a Chinese puzzle in itself. There are three key questions whose answers will guide you:

1. Is your data a sample or a population?
2. Does your scenario assume response knowledge and if so at what level?
3. If it is a sample then is it (approximately) a random sample?

By 'population' here we do not just mean the UK population (although that would be one example). For the purposes of statistics, a population is a complete set of objects or elements that share a particular characteristic of interest. For example, if your data are about all the members of the Bognor Regis Cycle club or all claimants of a certain benefit then these are a population (the word 'all' is the indicator here).

As we have discussed, response knowledge is a simple idea but it can be complex to apply. In some scenarios you may want to assume an intruder with *ad hoc* but full knowledge about a particular individual. For others you may want to consider a situation where the super-population (i.e. a larger set from which the population is drawn) is constrained. Perhaps I know that Bognor Regis cycle club members all have to live in Bognor Regis and own a bicycle; if I also know that you have those characteristics then I know that the probability of you being in the sample is considerably higher than I would estimate if I did not have that knowledge.

Response knowledge scenarios with microdata

First, consider the situation where your intruder has full response knowledge (to remind you, that is knowledge that a record corresponding to a particular known population unit is present in the microdata). This is the simplest case to assess. You need to identify how many unique combinations of the key variables you have in your dataset. This can be done simply using a spreadsheet or statistical package.

On the UKAN website you will find a set of CSV files that accompany this book which can be opened in a spreadsheet system or statistical software. These files contain example synthetic data that have been generated using some simple models, but which look like census data to give you something to practice on. Appendix B.1 gives instructions for calculating the number of uniques using Excel and Appendix

B.2 gives some syntax for use with statistical package SPSS. You can adapt either of these to your own data. In the example in the appendix, the key variables that we have used are age, sex, marital status, ethnic group, type of housing, tenure, number of cars, and whether the house has central heating. This corresponds to the type of things that neighbours might routinely know about one another. You might want to play around with other combinations of variables.

In the example data, using the "nosy neighbour" key we find that over 17% of the records are unique. What would such a result imply? Simply put, if our intruder has response knowledge for any of the individuals whose records are unique on those characteristics then they are at high risk of being re-identified. The only protection that these records have is the unreliable possibility of data divergence. Given that nearly 1 in 5 of the records in our dataset have this status we might decide that that is too high.

Faced with unique patterns in your population dataset what are your options? Essentially you have three choices: (i) give up now and do not release/share the data; (ii) proceed to apply disclosure control (see component 7 below); or (iii) if you still want to persist with your proposed release/share then you will need to carry out intruder testing (which we discuss shortly). If you go for option (ii) and you decide to apply data-focused, rather than environment-focused, disclosure control, then you will need to revisit this step once your data-focused control mechanisms have been applied, in order to reassess the risk.

Scenarios involving microdata without response knowledge

What if my file is not a population and my scenario analysis does not suggest response knowledge? Here you have a sample. There is a simple method known as data intrusion simulation (DIS)[88] which can help here. DIS provides a statistic which is straightforward to understand: *the probability of a correct match given a unique match*. In other words it tells you how likely it is that a match of auxiliary information against a record that is unique within the sample dataset is correct.[89] However, if you are looking at a strongly non-random sample then this step is a little trickier and you should consult an expert. Using the same data as the uniques test above, Appendix

[88] For a full technical description of the DIS method see Skinner and Elliot (2002). A brief explanation and some examples showing how it works are shown in Appendix C.

[89] A technical point; this method does assume that your sample is random, although in fact it is robust with respect of some degree of variance from random.

D.1 contains the instructions you need for implementing DIS in Excel and the SPSS syntax can be found in Appendix D.2.

The output of the DIS process is a measure of risk taken as the probability that a match against a unique in your dataset (on your selected key variable set) is correct. Essentially this takes account of the possibility that a unique record in a sample dataset may have a statistical twin in the population that is not represented in the sample.

For scenario keys of any complexity the outcome will not be zero risk. You knew this already of course, but quantifying risk immediately raises the question about how small a probability you should be aiming for. We cannot give you a single threshold because unfortunately there is no straightforward answer. Here instead in Table 3.2 is a set of rough guidelines for helping you think about your output. Given a particular quantitative output of the DIS process, we have mapped that onto a qualitative category on the assumption that the data are non-sensitive, and coupled with that an indication of the type of environmental solution that might be suitable. If the data are sensitive, then we need to shift down the table by one or even two categories, so that a DIS output of 0.03 should be treated as signalling a moderate risk (or even a high risk if the data are very sensitive), instead of the low risk it would signal on non-sensitive data. As ever in this field, context is all.

<0.001	Very low	
0.001-0.005	Low	Open data maximum
0.005-0.05	Low	End user licensed data maximum
0.05-0.1	Moderate	Restricted user licensed data maximum
0.1-0.2	High	On line remote access solutions maximum
>0.2	Very high	Highly controlled data centre solutions only

Table 3.2: classification of output from the DIS algorithm

To stress again, these are only for ball-park guidance but (assuming that your scenario analysis has been thorough) they should serve to indicate whether your overall level of risk is proportionate to your proposed solution.

If you have an unfavourable result at this stage, and you are out of your risk comfort zone (given the receiving data environment), what do you now do? The simplest solution at this stage is to apply aggregations to some of your key variables and/or sub-sample your data. Both of these will reduce the probabilities you have generated. Another alternative is to change the environment to one that is a lower

risk category. You may also revisit your use case. What constraints on data and environment are consistent with the needs of the use case?

One question that might arise is whether the above categorisations of DIS are a little on the conservative side. If the intruder only has a 1 in 5 chance of being correct, does that represent sufficient uncertainty regardless of the environment? If the risk was spread evenly across the file or if it were impossible for the intruder to pick out unusual records then that might be true but unfortunately, as noted earlier, some records are visibly more risky than others, vulnerable to fishing attacks or spontaneous recognition. This will be true even if you are broadly in your comfort zone.

Record level risk metrics

Conceptually, understanding disclosure risk at the record level is very simple: unusual combinations of values are high risk. Unfortunately, identifying all the risky combinations in a dataset is not straightforward and deciding what to do about them perhaps even less so. It might be at this point in the proceedings that you decide to call in the expert, but if you carry on there are some things that you can do that will at least have the happy side-effect of familiarising you with the data and their properties.

How does one define a risky record? There are many answers to this question and it is still an active research area. Yet focusing on the concept of uniqueness reveals two simple pragmatic principles:[90]

1. The more information you need to make a record unique (or to 'single it out') the less unusual it is.
2. The more information you need to make a record unique (or to 'single it out') the more likely that any match against it is prone to data divergence increasingly the likelihood of both false positive and false negative matches.

The first principle is primarily relevant to scenarios where there is sample data and no response knowledge. The second is relevant when either your data are population data or your scenario assumes response knowledge.

Start with principle 1. The basic idea is as follows. You have performed scenario analysis and generated a set of key variables. Say, for argument's sake, that you have

[90] For those that want to dig a bit deeper, there is some science underpinning this approach which is reported in Elliot et al (1998, 2002) and Haglin et al (2009).

eight of them.[91] Principle 1 says that if a record is unique in your data on, say, three of those variables it is more unusual than if you need values for all eight variables to make it unique. One way to think of this is that each time you add another variable to a key you divide the population or sample into smaller groups. Eventually everyone will be unique, so uniqueness itself is not such a big deal; the unusual people are those who are unique on a small number of categories.

The second proposition is in some ways simpler. Essentially each piece of information you have in your set of key variables carries with it the possibility of divergence. So each variable that you add to a key increases the probability of divergence for any match against that key. For sample data in scenarios without response knowledge, that has to be weighed against the informational gain from the additional variable. For population data or response knowledge scenarios the informational gain is irrelevant – a unique is a unique – but the impact of divergence is important and a sophisticated intruder will focus on individuals who are unique on a small number of variables.

Given these two principles, we can set out a rough and ready way of picking out unusual records. There is software available called SUDA[92] that can produce a more sophisticated version of this but understanding the principles first is still useful.

We will assume that you have an eight variable key, so you need to search for uniques on small subsets of those eight variables. We assume also that you have access to a statistics package.

1. First run all of the two variable cross tabulations. Do you have any uniques? If you do then identify the records that they belong to (filtering will do the trick here). These records are unusual enough to be noteworthy.
2. Now find the smallest non-unique cell in all of the two-way cross tabulations that you have run. Filter your datasets on that combination of values. Then run frequency tables on the remaining six variables. Are there any uniques in those frequency tables? If there are, then they will also be candidates of

[91] The astute reader may have noted that it is not just the number of variables that matters, but also the properties of those variables, most notably the number of categories, the skewness of the variable and correlations with other variables. However, these basic principles are sound and even this simplification will improve decision-making. We are considering steps that can be taken in-house, before calling in an expert consultant, and at some point it will be sensible to leave the complexities to them.

[92] Elliot and Manning (2003); available at: http://www.click2go.umip.com/i/software/suda.html. (accessed 30/5/2016).

interestingly unusual records (though probably not as unusual as the ones you identify in stage 1).

3. Repeat step 2 with next biggest non-unique cell in the two-way cross tables and continue repeating until you have reached a threshold (a cell size of 10 is a good rule of thumb).
4. Repeat steps 1-3 for each key variable set that you have.

This takes you as far as covering the 3-way interactions. In principle you can repeat the exercise with the 4-way interactions but that can involve a lot of output. Nevertheless, if your sample size in a response knowledge scenario is reasonably large then it might be important to do this.

Now you have a list of unusual records. What can you do with that? Well firstly, you can do a subjective assessment of the combination of values – do any of the combinations look unusual? You are likely to have knowledge of the general structure of the population, as you have a professional interest in the data, and that will undoubtedly help here. Perhaps present them to your colleagues to get a sanity check. Such subjective analysis is obviously not perfect and subject to all sorts of biases but nevertheless it can be informative (everyone would tend to agree 16 year old widowers are rare for example). If you have records that definitely appear unusual then you almost certainly need to take further action.

It is also important to consider how many records you have marked out as unusual. Is it a large portion of the size of your data file? If you have a relatively small number of records (relative to the file size), say less than 1%, then it might be possible to deal with them by techniques that involve distorting the data, which we consider in component 7 below. If the proportion is larger than that, then a more sensible approach is to carry out further aggregation and rerun the above analysis.

However, a cautionary tale will explain why it is also important to avoid knee-jerk reactions. A few years ago one of the authors was carrying out some work on behalf of a statistical agency identifying risky records within a longitudinal dataset, using the risk assessment software SUDA. The analysis threw up some odd patterns with some really high risk records. A bit of exploratory analysis revealed that these were records where the individuals had changed sexes several times in the space of a year or had reversed the ageing process. In other words they were the result of errors in the data. Arbitrary data errors will often lead to unusual looking records so not all unusual records are actually a risk and, more importantly, this sort of noise in the data generation processes does itself provide a handy side benefit of 'natural

protection' against intruders using fishing attacks (finding unusual records in the data and then attempting to find the equivalent unit in the population).

Penetration tests

There are essentially four stages to a penetration (pen) test: (i) data gathering; (ii) data preparation and harmonisation; (iii) the attack itself; and (iv) verification. The first stage tends to be the most resource intensive and the second and third require the most expertise. In general, external expert involvement will be helpful, even if you have the expertise yourself, to bring the perspective of an independent attacker.

Data gathering involves going out in to the world and gathering information on particular individuals. Exactly what that will look like will depend on the nature of the scenario that you are testing but would typically involve at least some searching of the Internet. The intruder test reported in Elliot et al (2016) gathered information on 100 individuals, taking about three person-months of effort. That test also included a second augmented attack using data purchased from the commercial data broker CACI.

A key point in this process is to decide whether one is assuming that the intruder has response knowledge or not – which will have been indicated by the scenario analysis. If so then the data holder will provide the matcher with a small sample of random formal identifiers (usually name and residential address), drawn from the dataset. If not then the simulated matcher will usually adopt the stance of finding unusual looking records in the dataset and attempting to find the corresponding individuals (the so called *fishing attack*).

Once the data gathering phase is complete then the data have to be harmonised with the target dataset. This will require work both across all the data, and at the level of individual records, as in all likelihood there will be several issues to address to achieve this. Gathered data will often be coded differently to the target data. For example you might have gathered information about somebody's job from social media, but how exactly would that be coded on the target dataset? There will be *data divergence* with the gathered information. For example the gathered and target data are unlikely to refer to the same set of time points so how likely is it that a given characteristic will have changed in the time differences and if so is that an important consideration? How confident are you in a piece of gathered information? For example Google Street View may show a motorcycle parked in the driveway of a target address. If you have a variable in your dataset indicating motorcycle

ownership, this is very tempting to adopt as a key piece of information, as it will be a highly skewed variable (most people do not own a motorcycle). But it may have belonged to a visitor, or the house might have changed ownership between the time of the Google visit and when the target dataset was created, or the bike might have been bought or sold in the interim. So when constructing your keys on a record-by-record basis you need to take into account all the information that you have gathered about a particular identity, but some of it should be flagged as less reliable at this preparatory stage so that it can be treated more cautiously at the attack stage.

Some scenarios simulate linkage between an identification dataset and a target dataset, rather than between gathered data and a target dataset. Here no data gathering is necessary but data harmonisation will still usually be necessary and issues of data divergence still critical, although the focus here will tend to be on the dataset as a whole rather than upon individual records.

The details of the attack stage will also depend on the nature of the data and the scenario. But typically it will involve attempting to link the information that you have gathered at stage 1 to your dataset. Usually this will involve a mixture of automated and manual processes. In essence you try to establish negative and positive evidence for links between your attack information and records in the dataset.

When you carry out the linkage you will quickly become aware that this is a non-exact science and the task is rarely as simple as dividing the potential matches into two piles. There is the matter of your confidence in the matches. This could simply be a subjective estimate of how likely you think it is that a match is a true match or it could involve a more quantitative approach. This will partly depend on what type of data intruder you are simulating. Is this an expert carrying out a demonstrative attack or simply the next door neighbour being nosy? Table 3.3 shows what an output from this process might look like.

We see from Table 3.3 that there are two individuals matched against record 42356 and that the individual 'Jane Indigo' is matched against two records. Here the matcher has been unable to distinguish cleanly between two possible matches against a record but is fairly confident that one of them is correct. It may be important to record these, because a real intruder may (again depending on the nature of the scenario) have options for *secondary differentiation* which are not available in the simulation. In other words, he or she may take close matches and engage using a different approach from the original data collection activity (for

example actually visiting a matched address and capturing further data by direct observation).

Name	Address	Record No	Confidence	Effective Confidence
Johnny Blue	10 Canterbury Gardens….	10985	95%	95%
Jamie Green	68 York Walk….	45678	95%	95%
William Pink	53 Winchester Lane….	42356	90%	60%
Fred Purple	39 Winchester Street….			30%
Archibald Black	68 Canterbury Walk….	671	85%	85%
Jane Indigo	23 Richmond Gardens….	37	80%	40%
		9985		40%
Patricia Vermilion	20 Winchester Drive….	70637	60%	60%
Wilma White	53 Lancaster Drive….	68920	50%	50%
Gertrude Gold	57 Privet Street….	35549	40%	40%
Brittany Magnolia	12 Acacia Walk….	22008	30%	30%
Petra Puce	75 Canterbury Street….	68680	30%	30%
Stephanie Red	11 Privet Drive….	81994	30%	30%
Simon Violet	136 Acacia Street….	91293	20%	20%
Estimated number of correct matches			7.05	

Table 3.3: An example output from a penetration test

A second point to note is that no match has 100% confidence associated with it. This reflects the reality that we can never be completely certain that we are correct. There is always a possibility that (i) the dataset contains data for a person who is highly similar to our target – their statistical twin – or that (ii) the assumption that our target is in the data is incorrect. It is worth noting in passing that this is the flipside of not being able to reduce the risk to zero.

Finally, once you have selected the matches, they need to be verified. This will often be carried out by a different person or organisation than the person doing the matching. If the matcher is carrying it out – at the risk of stating the obvious – they should only do this once they have decided upon their final list of matches.

In interpreting the results of a penetration test one needs to exercise some caution. Although the simulation will be a more direct analogue of what an actual intruder might do than with data analytical approaches, there are still differences which will impact on the results. Elliot et al (2016) list the following:

1. ***Ethical and legal constraints.*** Penetration tests are constrained ethically and legally; a real attack may not be.

2. ***Expertise variance.*** Typically the matcher will be an expert or at least skilled and knowledgeable about data. Even if they 'dumb down' their matching process in an effort to simulate a 'naive' intruder they will not be able to switch off their knowledge. This will particularly affect the estimation of match confidences.
3. ***Time available for data gathering.*** In order to get a picture of the risk across the whole dataset, pen tests usually consider multiple individuals. Resource constraints mean that the amount of time spent gathering information on each of those individuals will be limited. A real data intruder may be able to achieve their goal with just a single correct match and therefore may be able to focus attention on a specific individual.
4. ***Dataset specific results.*** Be careful about generalising any results to your data products and data situations in general.
5. ***Difficulties in simulating real response knowledge.*** A real data intruder with response knowledge might have ad hoc knowledge with respect of their target that it is hard to simulate through gathered data. If one wants to simulate such an attack, one would need to co-opt data subjects and members of their social network into the study. This is an interesting possibility, but to our knowledge no such study has ever been carried out and realistically would be too resource intensive for practical risk assessment.
6. ***Pen tests only give snapshots.*** The data environment is constantly changing and more specifically the availability of data that could be used to re-identify individuals is increasing. A pen test if done well may tell you a great deal about your risk now but that risk can and indeed will change.
7. ***Arbitrary variation of data divergence.*** Typically in these exercises one is gathering current data to carry out the simulated attack whereas the target data are past data. Temporal data divergence can markedly reduce the accuracy of matches so the degree of divergence between the data collection for the target dataset and the data gathering for the simulated attack will impact on the results.

Taking these considerations into account what sort of level of successful matching would one consider problematic? It is difficult to generalise this. But if you have produced a table like 3.3 and you see most of the high confidence matches are true matches then you have a problem and you need to rethink your data situation. But what if you have, say, a single correct match? The false positives are important here – are some of these high confidence matches? If so then the single correct match is swamped by false positives, in which case how could an intruder decide that that

match was a correct one? Remember they will not have the advantage of being able to verify!

One aspect to think about here is risk from the intruder's perspective – could claiming a match that turns out to be incorrect backfire on them? If so then they might well be cautious before making a claim. Another aspect to bring to the table in your thinking at this stage is the sensitivity of the data. If you think the impact of a correct match is high then your tolerance for a single correct match will be lower than if the expected impact is low.

Related to this is the importance of cross checking the correct match rate achieved against the rate estimated by the matcher. To derive the former, simply sum up the confidences (converted to proportions). So you can see in Table 3.3 the estimated number of correct matches is 7.05. If your number of correct matches varies significantly from this estimated figure then the matcher may have wrongly estimated their confidence level and it is worth considering calibrating the reported confidence levels so that the overall estimated number of matches is correct. The simplest method for doing this is to divide each confidence by the estimated number of matches and multiply by the number of correct matches achieved.

Of course a real data intruder might hit on a match which by chance happens to be correct, and they may not care or even know about nuances such as confidence levels. Although you have to think about such eventualities, you cannot build your data sharing practices around them – the correct place to deal with them is in your breaches policy, which we discuss in component 9.

A final question is what we assume the intruder knows about the disclosure control applied to the data. Nothing? The methods employed? The methods plus the parameters used? This will partly depend on the moment in the anonymisation process in which the penetration test is run, and the type of disclosure control that has been applied. If you have simply aggregated and deleted variables then we can assume that the intruder simply observes the effects of the control process. However, if data distortion has been applied then a sophisticated intruder will be able to use knowledge of the details of this if they are published. Note here that there is therefore an iterative relationship between this component and component 7, where disclosure controls are actually applied to the data.

Component 7: Identify the disclosure control processes that are relevant to your data situation.

Disclosure control processes essentially attend to either or both of the two elements of your data situation: the data and their environment. If your risk analysis in component 6 suggests that you need stronger controls then you have two (non-exclusive) choices:

1. Change the data (specification)
2. Reconfigure the data environment

In section 2.5.2 we described the various disclosure control techniques, and their pros and cons.

Changing the data

Usually one starts from a fairly fixed proposal of what the release/share environment will be, defined in components 1 and 4. It may be that this fixed idea has to change but initially one has to work on changing the data. The most common place to start is aggregation.

1. Keeping the use case in mind, can you lose detail on your key variables to reduce the measurable risk?
2. If your data situation is sensitive, can you remove or reduce detail on sensitive variables?

Often the answer is yes; you will lose some utility but not to the extent that the data lose most of their value.

Variables that tend to be a focus here are spatial and temporal ones – typically place of residence and age. The latter is particularly important if the data is about multi-member households. Other variables which can be considered are those with skewed distributions (where minority categories can be merged together). However, any variable that appears in your scenario keys should be considered.

At this point you should also consider producing a sample rather than releasing all of the data. Any level of sampling will reduce the risk, but sample fractions that one would normally consider range from between 1% and 50%. Most census and social surveys microdata products are released as samples at the bottom end of this range and these are generally regarded as high utility products, and so for release use cases give some serious consideration to this possibility.

One overarching advantage of metadata controls, such as aggregating scenario keys and sampling, is that you can easily rerun your risk measurements in order to see

what impact a particular aggregation has on the overall level of risk. Doing this with data distortion controls is more difficult.

In general, if it is possible to reduce the risk to an appropriate level through aggregation, variable deletion and sampling, then that should be the preferred approach. Applying data distortion controls affects the data utility in an unpredictable and non-transparent manner and leaves you with the difficult question about whether or not to release information about the distortion.

However, if you have done all that you think you might be able to do with metadata level controls and the risk is still too high, then you will have to move on to data distortions or reconfigure the environment. If the latter is not possible because the use dictates a particular environment, then distortion of the data is left as the only possibility.

Now you have to decide whether the distortion should be random or targeted. Random distortion in fact has relatively low impact on the risk –you will have to do quite a lot of distorting before you get a significant impact. Random distortion works by reducing the baseline confidence in any match. Targeted distortion potentially has a big impact on the disclosure risk. The point of targeting is to focus on the high risk records (those identified by your record-level risk metrics in component 6). So if you turn a sixteen year older widower into a sixteen year old single person then you have merged him into the crowd and the risk goes away. However, the big cost is that you alter variability and introduce bias. So our guidance is to do this only very sparingly.

The second issue is that once you have distorted the data then the standard risk metrics will no longer work. There are techniques for measuring post-distortion risk, but these are experimental and complicated to implement. So there are two options: (i) you add in distortion to pick up a small amount of residual targeted risk when you are quite close to your acceptable level anyway or (ii) you carry out a pen test.

In general, we would not advise using data distortion controls if you can avoid them and if you do you should consult an expert first.

Reconfiguring the environment

As described in chapter 2, reconfiguring the environment essentially involves controlling who has access, how they access the data and for what purposes. In some cases the environment is a fixed point of reference within the data situation – ‘we want to release an open version of this dataset’ or ‘we want to share these data with

organisation X' – in which case your anonymisation solutions will have to be data-focused and the environment will have been fixed as per components 1 and 4.

In other cases it is possible to achieve anonymisation, at least in part, through reconfiguring the environment. Options to consider are:

1. Allowing access only within your own secure environment.
2. Specifying the requisite level of security for the data.
3. Specifying that all analytical outputs must be checked and sanctioned by you before they are published.
4. Specifying the people who may access the data.

Placing or tightening controls on the environment will tend to have quite significant effects on the risk, often ruling out particular forms of attack, for example, and so if the data are sensitive they are certainly worth considering.

3.3 Impact Management

Much of what we have considered so far has framed risk management in terms of reducing the likelihood of an unintended disclosure happening, but it would be irresponsible not to prepare for the worst. Impact management puts in place a plan for reducing the impact of such an event should it happen.

Component 8: Identify your stakeholders and plan how you will communicate with them

Effective communication can help build trust and credibility, both of which are critical to difficult situations where you need to be heard, understood and believed. You will be better placed to manage the impact of a disclosure if you and your stakeholders have developed a good working relationship.

About your stakeholders

In component 4 we talked about the importance of communication and engagement with user groups. Your users are of course not the only group with an interest in your business or activity. Others who may be affected include data subjects, the general public, partner organisations, the media, funders and special interest groups.

Depending on the circumstances and the public interest in your data, many if not all the stakeholders just listed are likely to have an interest in your data, their use and reuse, whether confidentiality is a high priority in your organisation, and whether assurances of confidentiality are well-founded. However, what they would like to

hear about these topics may differ. For example, data subjects and the general public are likely to want to know the **what** of your processing activities, such as what data, in which environment(s). In contrast specialist interest groups and the media may also want to know about the **how** of your processing activities, such as how are they anonymised or how you determine an environment to be safe. The key is to engage (as well as communicate) with your stakeholders to determine what they would like to know about your processing activities, most obviously so that you can put your point of view across (and also, perhaps, adjust your practices in response to reasonable criticism), but also so that you understand their information needs immediately when you find you have to pick up the phone. You can do this in much the same way you engage with your user groups by, for example:

- **A web or mail survey:** You could develop a ten or twenty-minute survey to be delivered via your website or through a mail-out. Bear in mind you may need to tailor the survey to different stakeholder groups.
- **Going out and talking:** You may want to tailor the mode of discussion for particular target stakeholders, e.g. holding face-to-face meetings with funders, holding focus groups with representatives from the general public, etc.
- **A little research:** One way to identify concerns is to look at the type of FOI requests you and similar organisations in your sector receive. Identify common themes and whether particular stakeholders are associated with particular themes, e.g. a member of the public or the media.

Determining the next step once you know what your various stakeholders want to hear from you may or may not be straightforward. As we have already said, being open and transparent is always preferable but you may not be able to meet all your stakeholders' requests for information, either because they impact on your disclosure control or because they create their own confidentiality issues.

Communicating and engaging with stakeholders

Plan out how you will talk to and engage with your various stakeholders. Below is a list of pointers that you may wish to capture in your plan (it is not an exhaustive list).

Identify your key stakeholders

This is an obvious point but you need to make sure you capture all those likely to have a stake in your data processing activities; this might be a wide range of groups. Common stakeholders include those we have listed above, although this will be

dependent on your activities, your organisation, the sector you belong to etc. So, for example, stakeholders in the health sector in the UK will include groups such as the Department of Health, the local authority or council, hospital trusts, patients and patients' groups, service users, suppliers, funders, commissioning groups, quality assessors, special interest groups, community groups, the wider public health workforce, and the media.

Be clear about your aims and objectives in talking to your stakeholders.

This will help you ensure that your messages are clear and consistent. You may have multiple aims, such as: (i) to promote trust in your organisation's handling of data, (ii) to build relations with relevant specialist groups, and (iii) to promote awareness about your reuse of data for public benefit.

Your objectives should include details of how you will go about realising your aims. For example:

- If one of your aims is to *promote awareness in how you reuse data for public benefit* your objectives might be:
  Objective 1: Produce and publish case studies detailing how the reuse of x, y and z data has benefited the general public.
  Objective 2: Gather and publish testimonials on how the reuse of x, y and z data benefited particular groups.
  Objective 3: carry out and publish a Privacy Impact Assessment.

- If one of your aims is to *promote trust in your organisation's handling of data* your objectives might be:
  Objective 1: Produce and publish a clear statement outlining your commitment to data confidentiality.
  Objective 2: Produce a report on your data share and release activities.

Establish your key messages

This is critical to the effectiveness of your communications. Your key messages need to be clear and concise and address the concerns of your stakeholders.

About communication and public engagement activities

Your communication and public engagement activities should have clear timescales and goals to allow you to evaluate their effectiveness.

Examples of communications and engagement activities might include:

- **Press releases:** A concise press release can help you reach a large audience with little financial outlay.
- **Social media:** Regular and committed use of social networking, such as Facebook or Twitter, allows you to communicate immediately and in real time.
- **Actively maintain a website:** This will allow you to provide consistent messages over time, accessible to all (or most of) your stakeholders.[93]
- **Involvement and consultation activities:** Going out and meeting your stakeholders, holding focus groups, meetings, briefings and discussion forums etc., allows personal and face-to-face contacts to develop, which in many circumstances is more supportive of trust than a purely corporate outward face.

One final and very general point about communication is that by promoting trust, building relations and promoting any good works you will be helping to associate a positive view with your organisation's use of data. Promoting a positive view associated with your data products/organisation is important because you are operating in a complex global data environment over which you have limited control and bad news stories about data breaches and data security mishaps are all too frequent. If there has been a recent, widely-publicised data breach elsewhere in your sector, it may be that you, even though blameless, will be scrutinised by the media, or political campaigners, very closely for a period. If you have succeeded in establishing a positive view of your data practices, it may even mean that stakeholders who discover any problems with your data will be more likely to come to you quietly, enabling you to fix them, rather than immediately going public.

Component 9: Plan what happens next once you have shared or released the data

Having shared or released an anonymised dataset, do you need to do anything else in respect of those data? The simple answer is yes. It is our recommendation that you do not just release and forget about your data. Continuing advancements in IT capabilities, supporting ever-greater access to data and capacity for their analysis, and an ever increasing amount of available data, mean that there is always the potential for the data environment in which you have shared or released your data

[93] Examples where a lot of thought has been given to the key issues of public benefit and trust can be found at www.adrn.ac.uk and www.datasaveslives.eu/.

to change. So whilst your data may be considered safe at the time of its release this may not be the case in the medium term. This is a view also taken by the ICO:

> *'Means of identifying individuals that are feasible and cost-effective, and are therefore likely to be used, will change over time. If you decide that the data you hold does not allow the identification of individuals, you should review that decision regularly in light of new technology or security developments or changes to the public availability of certain records.'* (ICO Determining what is personal data. Version 1.1, 2012, page 9).

There are a number of measures you can take to monitor the data environment once you have shared or released your data. These measures should include (but are not limited to):

1. Keeping a register of all the data you have shared or released.
2. Comparing proposed share and release activities to past shares and releases to take account of the possibility of linkage between releases leading to a disclosure (as exemplified in section 2.3.4).
3. Be aware of changes in the data environment and how these may impact on your data. This means (i) keeping abreast of developments in new technologies and security that may affect your data situation by, for example, reading technology journals/blogs, watching relevant podcasts and/or attending relevant events; (ii) monitoring changes in the law or guidance on data sharing and dissemination by engaging with relevant organisations such as the ICO and UKAN and (iii) keeping track of current and new public data sources by, for example, reviewing the information available on the Internet and through more traditional sources such as public registers, local community records, estate agents' lists, professional registers, the library, etc.

If possible you should also keep track of how your data is used. If you are controlling access this is fairly straightforward; if you are releasing an open dataset then you might want to consider a process whereby users register their intended use before downloading. This type of information is invaluable later when you are considering the next release, developing its use case (component 4) and considering the risk and utility trade-offs in components 6 and 7.

If your organisation is large enough you may wish to appoint a Chief Data Officer to oversee these activities. Certainly you will need to ensure someone in your organisation takes responsibility for overseeing these measures.

Component 10: Plan what you will do if things go wrong

Sometimes, even when you follow all the recommended advice, things can go wrong. As identified in component 2 it is important that you have effective governance policies and procedures in place which essentially identify who does what, when and how, and generally support a culture of transparency. A natural extension of this is putting in place mechanisms that can help you deal with a disclosure in the rare event that one were to occur.

Ensure you have a robust audit trail

Being able to provide a clear audit trail taking into account all relevant anonymisation activities and processes will be crucial for the purpose of: (i) demonstrating that you have followed all correct procedures, and (ii) identifying where, if at all, in your processing activities you might need to make changes to prevent a similar occurrence. In practice this means keeping clear and up-to-date records of all your processing activities, detailing who did what, when and how. Some of this information can itself increase disclosure risk and thus these records may by default be internally facing. Not being transparent about the anonymisation process may, however, impact on utility and for this reason you may wish to provide a top level public narrative about your anonymisation processes.

Ensure you have a crisis management policy

A crisis management policy will ensure you deal effectively and efficiently with a data breach were one to occur. It should identify key roles and responsibilities and detail an action plan stating, step by step, the processes that should be followed in the event of a breach.

There are (at least) two key tasks within crisis management: managing the situation and communicating it to stakeholders. These tasks, if taken on by more than one person, require close cooperation from the start right through to the post-breach review.

Ensure you have adequately trained staff.

You should ensure that all staff involved in your data processing activities are suitably skilled and experienced for the tasks they undertake and that they understand their responsibilities.

You will in all likelihood need to conduct training to ensure staff are kept up-to-date with relevant anonymisation issues. This might take the form of:

- In-house training on the principles and procedures of your data processing activities.
- External training on core factors such as anonymisation issues and techniques, data security, data protection law etc.

Other ways to support the safe handling of data might include:

- Organising regular team meeting/briefings to look at anonymisation issues such as 'what are my responsibilities under a Data Share Agreement when processing data from another source?'
- Implementing a staff non-disclosure agreement to provide clear guidance to staff about their data confidentiality responsibilities inside and outside of their workplace and when employment at your organisation ceases.

Managing the situation

Set out a plan for managing the situation. The types of activities you will need to cover are outlined in steps 1 to 6 below. By establishing step-by-step what you will need to do will help you both better manage the situation and avoid having to make decisions in haste.

In your plan you should identify the person who will take overall responsibility for managing the situation. You should also include a clear description of their responsibilities.

In the event of a data breach your staff will need to know their roles and responsibilities. Your plan should make these clear. For example, when a member of staff first becomes aware of a breach what should they do? Who should they contact and how? What should they do if the person identified as the first point of contact is not immediately available?

Communicating the situation

Within your crisis management plan you will need to detail a strategy for communicating with key stakeholders, especially those who may potentially be directly affected by the breach, the ICO, the media and other interested parties. You should identify a spokesperson to represent you/your organisation to ensure your messages about the breach and your responses to it are clear and consistent. Transparency is always preferable but you will probably need time to get all the key information together so you may need an initial holding response to stakeholders such as '*we are investigating the matter*'. Nevertheless, it is important that you are

more concrete and on the record about what you are doing as early as possible in the process.

Steps in a crisis management plan

More widely, the key point is that everyone in your organisation should know what your strategy is and their role in it. A plan for managing a data breach might include the following steps:[94]

Step 1: Respond swiftly

Include in the plan the first series of actions for a range of possible relevant situations and how they might be undertaken. For example, in the event of a breach relating to datasets published on (our) website immediately take the dataset down from the website.

Step 2: Assess the impact

Include in the plan how the potential impact might be assessed and recorded. The key questions here would be:

- Can you guesstimate the potential for other copies of the data being in existence – e.g. from knowledge of users, website traffic?
- What is the nature of the breach?
- Are the data sensitive?
- Is anyone, and if so how many are, likely to be affected by the breach?
- What is the nature of the harm likely to be experienced?

Step 3: Put measures in place to limit the impact

Include a feedback loop so that once step 2 is completed you can reconsider if any further interim action can be taken. Think through the types of further action that might be required and plan how you would deliver them.

Step 4: Notify the appropriate people.

Include in the plan details about who should be notified about the breach, how and within what timeframe.[95]

---

[94] Please also see the ICO's guidance on managing a data breach; UK: Information Commissioner's Office (2012b).

[95] For further information see UK: Information Commissioner's Office (2016). We note that under the new EU regulations notification will be mandatory; European Commission (2015).

Step 5: Penalties

Include in the plan details about any penalties associated with behaviours indirectly or directly leading to a breach. Make sure identified penalties are fair, consistent and enforceable.

Step 6: Review the breach and your handling of it

The aim here is to learn lessons from the event and put procedures in place to prevent a further occurrence. You should stipulate who will undertake the review and within what time frame.

Ensure you undertake a periodic review of your processing activities

A review process is likely to be most effective if it is undertaken periodically and not just when a crisis occurs. You should stipulate who is responsible for the review, when and how it will be undertaken and within what time frame. For this you might want to develop your own standardised form that captures your data processing activities and the criteria against which they will be assessed.

3.4 Closing remarks

In this chapter, we have described the anonymisation decision making framework as a practical tool for dealing with your data situation. As we said at the outset, the framework is not a simple checklist system but does provide structure which can reduce the complexity of the process of anonymising your dataset before you release or share it. Each of the components in the framework will require thought and planning to implement, but with appropriate resourcing can turn data sharing from a confusing and daunting process that puts you under pressure, into a practical and perhaps even exciting possibility for optimising the utility of your data.

3.4.1 Further reading

For the reader who is interested in going deeper into any of the topics that the framework covers there is a wealth of material available both in print and online. Around the particular issues to do with the technicalities of disclosure risk assessment and control there are several technical primers. The easiest of them is probably Duncan et al (2011) which starts with three conceptual chapters only a little beyond the material presented here before launching into more technical material. The most comprehensive treatment of orthodox data focused disclosure control can be found in Hundepool et al (2012). A good source for finding out about the state of the art in disclosure control is the *Privacy in Statistical Databases* series which is edited

by Josep Domingo-Ferrer and colleagues and published every two years. The last edition was published in 2014.

For treatments of the end to end anonymisation problem that particularly focus on health data we would recommend the reader looks at the work of Khaled El-Amam and colleagues, particularly the 2013 edited collection entitled *Risky Business* and two recent authored books *Anonymizing Health Data* (2014) and *Guide to the De-Identification of Personal Health Information* (2013). These are primarily aimed at the North American market but like our own offering here there is much that is transferable to other jurisdictions.

Discussions of the ethical and legal issues surrounding anonymisation and data sharing are numerous. We do particularly recommend Helen Nissenbaum's (2010) book on contextual integrity. Also of note is her recent (2015) collaboration with Finn Brunton called *Obfuscation* which could be read as a call for data subjects to anonymise their own data and possibly serves as a warning of what is likely to happen if data holders do not get their privacy practices into a better space.

On the specific issue of consent, this is very much an area of open debate. We refer the reader to articles by Singleton and Wadworth (2006), Iversen et al (2006) and Haynes et al (2007) for general discussions about the issue. More recently, Cruise et al (2015) discuss consent issues related to data linkage and Hallinan and Friedewald (2015) raise the import issue of whether consent can ever truly be informed – and the relevance of this to the new EU general data protection regulation.

If you are considering how this all fits in with the new world of big data, Van Den Hoven et al (2015) is a good place to start. Julia Lane and colleagues' (2014) edited collection is also very good for some serious thinking about the direction of travel. Other recent perspectives are provided by: Crawford and Schultz (2014), Boyd and Crawford (2012), Szonsgott et al (2012), Rubenstein (2013), Richards and King (2013), Narayanan et al (2016), and Matzner et al (2016). The volume, velocity and variety of opinions, perspectives and new ideas in this area mirrors the properties of big data itself. Suffice to say if you are dipping into this literature expect to come away with more questions than answers.

3.4.2 Next steps for the framework

We regard this framework as an organic open document. The data environment is constantly changing and new forms of data are appearing all the time. Therefore, we will be reviewing the framework and revising it on a regular basis. Please do use the

feedback form available from the UKAN website to provide input into this ongoing development.

We will also be developing new case studies that explicitly use the framework and if you are interested in working with us to develop such a case study using your data situation then please do get in touch.

References

ARRINGTON, M. (2006) *AOL proudly releases massive amounts of user search data,* TechCrunch, available at: http://tinyurl.com/AOL-SEARCH-BREACH [accessed 30/5/2016].

ATOKAR (2014) *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset,* available at: http://tinyurl.com/NYC-TAXI-BREACH [accessed 30/5/2016].

BATESON, N. (1984) *Data Construction in Social Surveys*. London: George Allen and Unwin.

BOURNE, I. (2015) *Personal correspondence.*

BOYD, D. & CRAWFORD, K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon; *Information, communication & society*, 15(5): 662-679, available at: http://dx.doi.org/10.1080/1369118X.2012.678878 [accessed: 30/5/2016].

BRUNTON, F. & NISSENBAUM, H. (2015) *Obfuscation: A User's Guide for Privacy and Protest*. Cambridge: MIT Press.

CNN MONEY (2010) *5 data breaches: From embarrassing to deadly*, available at: http://tinyurl.com/CNN-BREACHES/ [accessed: 30/5/2016].

CRAWFORD, K. & SCHULTZ. J. (2014) Big data and due process: Toward a framework to redress predictive privacy harms; *BCL Rev,* 55(1): 93, available at: http://tinyurl.com/BD-HARMS [accessed 30/5/16].

CRUISE, S. M., PATTERSON, L., CARDWELL, C. R., & O'REILLY, D. (2015) Large panel-survey data demonstrated country-level and ethnic minority variation in consent for health record linkage; *Journal of clinical epidemiology*, 68(6): 684-692, available at: http://tinyurl.com/LINK-CONSENT [accessed 30/5/16].

DE MONTJOYE, Y. A., RADAELLI, L., & SINGH, V. K. (2015) Unique in the shopping mall: On the reidentifiability of credit card metadata; *Science*, 347(6221): 536-539, available at: http://tinyurl.com/UNIQ-CC [accessed 30/5/16].

DOBRA, A. & FIENBERG, S. E. (2000) Bounds for cell entries in contingency tables given marginal totals and decomposable graphs; in *Proceedings of the National Academy of Sciences*, 97(22): 11885-11892, available at: http://tinyurl.com/BNDS-DECOM [accessed 30/5/16].

DOBRA, A. & FIENBERG, S. E. (2001) Bounds for cell entries in contingency tables induced by fixed marginal totals; *Statistical Journal of the United Nations Economic Commission for Europe,* 18(4): 363-371, available at: http://tinyurl.com/BNDS-MARGINAL [accessed 30/5/16].

DOMINGO-FERRER, J. & TORRA, V. (2008) A critique of k-anonymity and some of its enhancements; In *3rd Intl. Conference on Availability, Reliability and Security (ARES 2008)*, Los Alamitos CA: IEEE Computer Society, 2008: 990-993, DOI: 10.1109/ARES.2008.97.

DOMINGO-FERRER, J, SÁNCHEZ, D., & SORIA-COMAS, J. (2016) *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections;* Synthesis Lectures on Information Security, Privacy, & Trust 15: Morgan & Claypool, DOI: 10.2200/S00690ED1V01Y201512SPT015.

DOMINGO-FERRER, J. (ed.) (2014) *Privacy in Statistical Databases*. Lecture Notes in Computer Science 8744, Heidelberg: Springer-Verlag.

DOYLE, P., LANE, J. I., THEEUWES, J. M. & ZAYATZ, L. V. (eds.) (2001) *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier, 135-166.

DUNCAN, G. T., FIENBERG, S. E., KRISHNAN, R., PADMAN, R., & ROEHRIG, S. F. (2001) Disclosure limitation methods and information loss for tabular data; In Doyle, P., Lane, J. I., Theeuwes, J. M., & Zayatz, L. V. (eds.) *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier, 135-166.

DUNCAN, G. T., ELLIOT, M. J., & SALAZAR-GONZÁLEZ, J. J. (2011) *Statistical Confidentiality*. New York: Springer.

DUNCAN, G. & LAMBERT, D. (1989) The risk of disclosure for microdata; *Journal of Business & Economic Statistics*, 7(2): 207-217, DOI: 10.1080/07350015.1989.10509729.

DWORK, C., MCSHERRY, F., NISSIM, K., & SMITH, A. (2006) Calibrating noise to sensitivity in private data analysis; In Halevi, S. & Rabin, T. (eds.). *Theory of Cryptography,* Berlin Heidelberg: Springer-Verlag. pp. 265-284.

ELAMIR, E.A. & SKINNER, C. (2006) Record level measures of disclosure risk for survey microdata; Jou*rnal of Official Statistics*, 22(3): 525, available at: http://tinyurl.com/REC-RISK [accessed 30/5/16].

ELKINGTON, J. (1997) *Cannibals with Forks: the Triple Bottom Line of 21st Century Business*. Gabriola Island, BC Canada: New Society Publishers.

EL EMAM, K. (2013) *Guide to the De-Identification of Personal Health Information*. Boca Raton, Florida: Auerbach Publications (CRC Press).

EL EMAM, K. (ed.) (2013) *Risky Business: Sharing Health Data while Protecting Privacy.* Bloomington, Indiana: Trafford Publishing.

EL EMAM, K. & ARBUCKLE L. (2014) *Anonymizing Health Data 2nd Edition*. Sebastapol, California: O'Reilly media.

ELLIOT, M. J. (1996) Attacks on Confidentiality Using the Samples of Anonymised Records; In *Proceedings of the Third International Seminar on Statistical Confidentiality*. Bled, Slovenia, October 1996. Ljubljana: Statistics Slovenia-Eurostat.

ELLIOT, M. J. (2000) DIS: A new approach to the measurement of statistical disclosure risk; *Risk Management*, 2: 39-48, DOI:10.1057/palgrave.rm.8240067.

ELLIOT, M. J. (2001) Advances in data intrusion simulation: A vision for the future of data release; *Statistical Journal of the United Nations Economic Commission for Europe*, 18(4): 383-391.

ELLIOT, M. J. & DALE A. (1998) *Disclosure risk for microdata;* report to the European Union ESP/204 62 (1998): 361-372.

ELLIOT, M. J. & DALE, A. (1999) Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk; *Netherlands Official Statistics*, Spring 1999: 6-10, avilable at: http://tinyurl.com/ATTACK-SCENARIO [accessed 30/5/16].

ELLIOT, M. J., DIBBEN, C., GOWANS, H., MACKEY, E., LIGHTFOOT, D., O'HARA, K., & PURDAM, K. (2015) Functional Anonymisation: The crucial role of the data environment in determining the classification of data as (non-) personal; CMIST work paper 2015-2 available at http://tinyurl.com/FUNC-ANON [accessed 27/5/2016].

ELLIOT, M. J. & MACKEY, E. (2014) The Social Data Environment; In O'Hara, K., David, S. L., de Roure, D. Nguyen, C. M-H. (eds.), *Digital Enlightenment Yearbook,* pp. 253-263.

ELLIOT, M. J., MACKEY, E., O'SHEA S., TUDOR, C. & SPICER, K. (2016) Open Data or End User License: A Penetration Test; *Journal of Official Statistics*, 32(2): 329–348, DOI: 10.1515/JOS-2016-0019.

ELLIOT, M., J., MACKEY, E. & PURDAM, K. (2011) *Formalizing the Selection of Key Variables in Disclosure Risk Assessment*, 58th Congress of the International Statistical Institute, Aug 2011, Dublin, Ireland.

ELLIOT, M. J. & MANNING, A. M., (2003) *SUDA: A software tool for use with statistical disclosure control for microdata,* Manchester: UMIP, available at http://www.click2go.umip.com/i/software/suda.html [accessed 30/5/2016].

ELLIOT, M. J., MANNING, A. M. & FORD, R. W. (2002) A computational algorithm for handling the special uniques problem; *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 493-509, DOI: 10.1142/S0218488502001600.

ELLIOT, M. J., SKINNER, C. J. & DALE, A. (1998) Special Uniques, Random Uniques and Sticky Populations: some counterintuitive effects of geographical detail

on disclosure risk; *Research in Official Statistics*, *1*(2): 53-67. DOI: 10.2307/3867923.

EU: ARTICLE 29 DATA PROTECTION WORKING PARTY (2007) Opinion 4/2007 on the concept of personal data (Adopted 20th June 2007), 01248/07/EN WP136, available at: http://tinyurl.com/WP29-PERS-DATA [accessed 30/5/2016].

EU: ARTICLE 29 DATA PROTECTION WORKING PARTY (2010) Opinion 1/2010 on the concept of 'controller' and 'processor', (Adopted 16th February 2010). 00264/10/EN WP169, available at: http://tinyurl.com/WP29-CONT-PROC [accessed 31/5/2016].

EU: *Directive 95/46/EC - The Data Protection Directive* (1995) available at: http://tinyurl.com/EU-DPD95 [accessed 30/5/2016].

EUROPEAN COMISSION (2015) *Questions and Answers - Data protection reform* available at: http://tinyurl.com/EC-DP-MEMO [accessed 30/5/2016].

FIENBERG, S.E. (2005) Confidentiality and Disclosure Limitation; In Kempf-Leonard K. (ed.), *Encyclopaedia of Social Measurement volume 3*. New York: Elsiever.

GUTWIRTH, S., LEENES, R. & DE HERT, P. (eds.) (2016*) Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection* (Vol. 24). Heidelberg: Springer.

HAGLIN, D. J., MAYES, K. R., MANNING, A. M., FEO, J., GURD, J. R., ELLIOT, M. J. & KEANE, J. A. (2009) Factors affecting the performance of parallel mining of minimal unique item sets on diverse architectures; *Concurrency and Computation: Practice and Experience*, 21(9): 1131-1158, DOI: 10.1002/cpe.1379.

HALLINAN, D. & FRIEDEWALD, M. (2015) Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation?; *Life Sciences, Society and Policy*, 11(1): 1-36, available at: http://tinyurl.com/j3mrj36 [accessed 31/5/16].

HALEVI, S. & RABIN, T. (eds.) (2006) *Theory of Cryptography.* Berlin Heidelberg: Springer-Verlag.

HAYNES, C. L., COOK, G. A., & JONES, M. A. (2007) Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease register; *Journal of Medical Ethics*, 33(5): 302–307, DOI:10.1136/jme.2006.016907.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. & DE WOLF, P. P. (2012) *Statistical Disclosure Control*. London: John Wiley & Sons.

IVERSEN, A., LIDDELL, K., FEAR, N., HOTOPF, M. & WESSELY, S. (2006) Consent, confidentiality, and the data protection act; *British Medical Journal*, *332*(7534): 165-169, DOI: 10.1136/bmj.332.7534.165.

KEMPF-LEONARD K. (ed.) (2005) *Encyclopaedia of Social Measurement*. New York: Elsiever.

LANE, J., STODDEN, V., BENDER, S. & NISSENBAUM, H. (Eds.) (2014) *Privacy, Big Data, and the Public Good.* Cambridge: Cambridge University Press.

MACKEY, E. (2009) *A framework for Understanding Statistical Disclosure Control processes;* PhD Thesis, The University of Manchester. Manchester: University of Manchester.

MACKEY, E. & ELLIOT, M. J. (2013) Understanding the Data Environment; *XRDS: Crossroads,* 20 (1): 37-39.

MATZNER, T., MASUR, P.K., OCHS, C.,& VON PAPE, T. (2016) Do-It-Yourself Data Protection—Empowerment or Burden?; In Gutwirth, S., Leenes, R., & De Hert, P. (eds.) *Data Protection on the Move*, pp. 357-385. Heidelberg Springer.

MCCULLAGH, K. (2007) Data sensitivity: proposals for resolving the conundrum; *Journal of International Commercial Law and Technology*, 2 (4): 190-201, available at: http://tinyurl.com/z874zjp [accessed 30/5/2016].

MYSCOCIETY (2016) *What do they know.* London: MySociety[online], available at https://www.whatdotheyknow.com/ [accessed 30/5/2016].

NARAYANAN, A., HUEY, J., & FELTEN. E. W. (2016) A Precautionary Approach to Big Data Privacy; In Gutwirth, S., Leenes, R., & De Hert, P. (eds.) *Data Protection on the Move,* pp. 357-385. Heidelberg Springer.

NISSENBAUM, H. (2004) Privacy as contextual integrity; *Washington Law Review*, 79 (119): 101-139, available at: http://tinyurl.com/j8xut58 [accessed 30/5/2016].

NISSENBAUM, H. (2010) *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Palo Alto, CA: Stanford University Press.

OFFICE FOR NATIONAL STATISTICS (2016a) *Virtual Microdata Laboratory (VML)* [Online] available at: http://tinyurl.com/ONS-VML [accessed 30/5/2016].

OFFICE FOR NATIONAL STATISTICS (2016b) *About the Longitudinal Study (LS)* [Online] available at: http://tinyurl.com/ONS-LS [accessed 30/5/2016].

OHM, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization; *UCLA Law Review*, 57: 1701-1777, available at: http://www.uclalawreview.org/pdf/57-6-3.pdf [accessed 30/5/2016].

O'KEEFE C.M. & CONNOLLY, C. (2010) Privacy and the use of health data for research; *Med J Australia* 193 (2010), pp 537-541, available at: http://tinyurl.com/zxgnhvq [accessed 30/5/2016].

O'KEEFE, C.M., GOULD P. & CHURCHES, T. (2014) Comparison of two remote access systems recently developed and implemented in Australia; In Domingo-Ferrer J. (Ed.), *Privacy in Statistical Databases* 2014, LNCS 8744, pp 299-311.

O'KEEFE, C.M., WESTCOTT, M., CHURCHES, T., ICKOWICZ A. & O'SULLIVAN, M. (2013) Protecting Confidentiality in Statistical Analysis Outputs from a Virtual Data Centre; In *UNECE/Eurostat Work Session on Statistical Data Confidentiality*. Ottawa, Oct 2013, available at: bit.ly/1RQGgej [accessed 18/2/16].

PURDAM, K. & ELLIOT, M. J. (2007) A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records; *Environment and Planning A*, *39*(5): 1101-1118, DOI: 10.1068/a38335.

REITER, J.P. (2005) Estimating Risks of Identification Disclosure in Microdata; *Journal of the American Statistical Association* 100(472): 1103-1112. DOI: 10.1198/016214505000000619.

RICHARDS, N. M. & KING, J. H. (2013) Three Paradoxes of Big Data; *Stanford Law Review Online* 41 (2013). Available at: http://ssrn.com/abstract=2325537 [accessed 28/5/16].

RUBINSTEIN, I. S. (2013) Big data: the end of privacy or a new beginning?; *International Data Privacy Law,* 3(2): 74-87. Available at: http://tinyurl.com/q3wvd53 [accessed 28/5/16].

SAMARATI, P. (2001) Protecting respondents' identities in microdata release; *IEEE Transactions on Knowledge and Data Engineering,* 13(6): 1010–1027.

SAMARATI, P. & SWEENEY, L. (1998) Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Washington: SRI International, available at: http://tinyurl.com/sam-swe-kanon [accessed 28/5/16].

SÁNCHEZ, D., MARTÍNEZ, S. & DOMINGO-FERRER, J. (2016) Comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata'; *Science* 351 (6279): 1274.

SARATHY, R. & MURALIDHAR, K. (2011) Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data; *Transactions on Data Privacy*, *4*(1): 1-17, available at: http://www.tdp.cat/issues11/tdp.a040a10.pdf [accessed 28/5/16].

SIMITIS, S. (1999*) Revisiting sensitive data*, available at http://tinyurl.com/j4w35bp

[accessed 19/6/2016].

SINGAPORE: *Personal Data Protection Act* (2012) Singapore: Singapore Statutes Online, available at: http://tinyurl.com/SING-DPA [accessed 30/5/16].

SINGLETON, P. & WADSWORTH, M. (2006) Confidentiality and consent in medical research: Consent for the use of personal medical data in research; *British Medical Journal*, 333(7561): 255, available at: http://tinyurl.com/jc8f3zo [accessed 30/5/16].

STATISTICS CANADA (2015) *What prevents a researcher from removing data from an RDC?*, available at: http://www.statcan.gc.ca/eng/rdc/faq#a8 [accessed 30/5/16].

SKINNER, C. J. & ELLIOT, M. J. (2002) A measure of disclosure risk for microdata; *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64(4): 855-867, DOI: 10.1111/1467-9868.00365.

SMITH, D. & ELLIOT, M. (2008) A Measure of Disclosure Risk for Tables of Counts; *Transactions on Data Privacy*, 1(1): 34-52, available at: http://www.tdp.cat/issues/tdp.a003a08.pdf [accessed 30/5/16].

SZONGOTT, C., HENNE, B. & VON VOIGT, G. (2012). Big data privacy issues in public social media; In *6th IEEE International Conference on Digital Ecosystems Technologies (DEST)*, pp. 1-6. Campione d'Italia, Italy, 18 – 20 June 2012, IEEE.

THOMPSON, G., BROADFOOT, S. & ELAZAR, D. (2013) Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics; In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Ottawa, Canada, 28–30 October 2013, available at: bit.ly/1PTippv [accessed 17/2/16].

TVERSKY, A. & KAHNEMAN, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157): 1124-1131, available at: http://tinyurl.com/o886vjm [accessed 30/5/16].

UK: *Census (Confidentiality) Act (1991)* London: The Stationery Office [Online], available at: http://tinyurl.com/UK-CENSUS-ACT [Accessed: 30/5/2016].

UK: *Commissioners for Revenue and Customs Act* (2005) London: The Stationery Office [Online], available at: http://tinyurl.com/UK-CRCA [Accessed: 30/5/2016].

UK: *Data Protection Act* (1998) London: The Stationery Office [Online], available at: http://tinyurl.com/UK-DPA98 [Accessed: 20/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2011) *Data sharing code of practice*, available at http://tinyurl.com/ICO-SHARE [accessed 25/5/2016].

WELLCOME TRUST (2016) *Public attitudes to commercial access to health data*, available at http://tinyurl.com/zl2es5b [accessed 19/6/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2012a) *Anonymisation: managing data protection risk code of practice,* available at http://tinyurl.com/ICO-ANON [accessed 25/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2012b) *Guidance on data security breach management,* available at http://tinyurl.com/ICO-BREACHES [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2012c) *Determining what is personal data. Version 1.1,* available at http://tinyurl.com/ICO-WHATISPD [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2014a) *Data controllers and data processors: what the difference is and what the governance implications are,* available at http://tinyurl.com/ICO-CONT-PROC [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2014b) *Conducting privacy impact assessments code of practice,* available at http://tinyurl.com/ICO-PIA2 [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2016) *Guide to Data Protection version 2.4,* available at http://tinyurl.com/ICO-DPG2-4 [accessed 30/5/2016].

UK: NATIONAL ARCHIVES, (2016) Open Government License, National Archives: Kew, UK [Online], available at http://tinyurl.com/NA-OGLv3 [accessed 30/5/2016].

UK: Statistics and Registration Services Act (2007) London: The Stationery Office [Online], available at http://tinyurl.com/UK-SRSA [Accessed: 30/5/2016].

VAN DEN HOVEN, J., HELBING, D., PEDRESCHI, D., DOMINGO-FERRER, J., GIANOTTI, F. & CHRISTEN, M. (2012) FuturICT – The road towards ethical ICT; *The European Physical Journal-Special Topics*, 214:153-181, available at http://tinyurl.com/ETHICAL-ICT [accessed 30/5/16].

WILLENBORG, L. & DE WAAL, T. (2001) *Elements of Disclosure Control*. Springer: New York.

Glossary of Terms

Additivity: A feature of **tables of counts** where the column and row totals are exact sums of the columns and rows they correspond to. **Rounding** and other forms of data distortion can violate additivity.

Anonymisation: This is a complex process that transforms **identifiable data** into non-identifiable (anonymous) data. This usually requires that identifiers be removed, obscured, aggregated and/or altered in some way. It may also involve restrictions on the **data environment**.

Analysis Server: A system – often virtual – where data users do not access data directly but instead submit analytical requests which are run (usually automatically) and then the users provided with the analytical output. That output may be checked for disclosiveness or the system maybe set up so as to only allow a restricted range of requests known to produce only safe output.

Analytical Completeness: A measure of the capacity of a dataset (in terms of variables, variable codings and sample size) to deliver a given analysis. Completeness is measured relative to some reference dataset usually the dataset before disclosure control has been applied.

Analytical Validity: A measure of whether a dataset produces the same result for a given analysis as a reference dataset (usually the dataset before disclosure control has been applied).

Attribution: This is the process of associating a particular piece of data with a particular **population unit** (person, household business or other entity). Note that attribution can happen with **re-identification** (if for example all members of a group share a common attribute).

Barnardisation: A form of **noise addition** for aggregate tables of counts where small numbers (usually -1, 0 and +1) are added to each cell.

Confidence: a measure, often subjective, of the certainty with which an intruder (or matcher within a penetration test) believes that a match between a population unit and a data unit is correct.

Confidentiality: The protection of the data/information from unwanted disclosure. With **personal data** this concerns the disclosure of identified or identifiable information.

Data controller: An entity that makes decisions about the processing of some data. Note that being a data controller is not a singular role (in the manner of say a Caldecott guardian) but a relationship between an entity and the data they are controller of.

Data environment: This is an explanatory concept in the realm of data privacy. It is best understood as the context for any item of data.

Data distortion controls: Any method of disclosure control which controls disclosure risk by manipulating the variable values at the level of individual **data units**.

Data divergence: This represents the differences between two datasets (data-data divergence) or between a single dataset and reality (data-world divergence). Sources of data divergence include: data ageing, response errors, mode of collection, coding or data entry errors, differences in coding and the effect of disclosure control.

Data intruder: A data user who attempts to disclose information about a data subject through identification and/or attribution (see statistical disclosure). Intruders may be motivated by a wish to discredit or otherwise harm the organisation disseminating the data, to gain notoriety or publicity, or to gain profitable knowledge about particular data subjects. The term also encompasses inadvertent intruders, who may spontaneously recognise individual cases within a dataset. Data intruders are sometimes referred to as ***attackers*, *snoopers* or *adversaries*.**

Data processor: An entity that processes personal data on behalf of a data controller but does not make decisions about that processing.

Data protection: This refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into data subjects' privacy caused by the collection, storage and sharing of data.

Data release: Any process of data dissemination where the **data controller** no longer directly controls who has access to the data. This ranges from general **licensing** arrangements, such as end user licensing where access is available to certain classes of people for certain purposes, through to fully **open data** where access is unrestricted.

Dataset: Any collection of data about a defined set of entities. Normally employed to mean data where data units are distinguishable (i.e. not summary statistics).

Data share: A dynamic data situation where the data controller has made a decision to allow a fixed set of entities access to a given dataset.

Data Situation: The relationship between some data and their environment.

Data Situation audit: The initial stage of the anonymisation decision making framework that clarifies the nature of the data situation and the elements that require further analysis.

Data subject: An individual to whom a particular piece of data relates.

Data swapping: A method of statistical disclosure control which involves swapping the values of a **key variable** (most often geography) between records which are similar on some other set of variables (often household composition).

Data unit: A case in a dataset; a set of data about a single population unit.

Data user: An entity (person or organisation) that processes data. In the context of anonymisation it is usually employed to mean that the data are non-personal and therefore users are not data controllers or data processors.

Data utility: A term describing the value of a given data release as an analytical resource - the key issue being whether the data represent whatever it is they are supposed to represent. **Disclosure control** methods can have an adverse effect on data utility. Ideally, the goal of any disclosure control regime should be to maximise data utility whilst minimising disclosure risk. In practice disclosure control decisions are a trade-off between utility and **disclosure risk**.

De-identification: The removal or masking of formal identifiers within a dataset.

Differencing: A re-identification attack whereby two different and overlapping codings for a variable (usually geography but in principal are variable) are overlain leading to intersecting categories which contain small numbers of cases.

Disclosure control methods: These are a set of methods for reducing the risk of disclosure, such methods are usually based on restricting the amount of, or modifying the, data released.

Disclosive data: Data are considered to be disclosive when they allow data subjects to be identified, (either directly or indirectly) and/or when they allow information about data subjects to be revealed. Data can be disclosive without any actual disclosures having happened.

Disclosure risk: This is expressed as the probability that an **intruder** identifies and/or reveals new information about at least one data subject in the disseminated data. Because anonymisation is difficult and has to be balanced against **data utility**, the risk that a disclosure will happen will never be zero. In other words there will be a risk of disclosure present in all useful anonymised data.

Direct identifier: Any data item that, on its own, could uniquely identify an individual case. It is sometimes referred to as a direct identifier, examples of which include a data subject's name, address and unique reference numbers, e.g. their social security number or National Health Service number.

Dynamic data situation: A data situation where data is being moved from one **data environment** to another.

Equivalence class: A set of data units that are identical on a given set of variables.

Equivalence class structure: A frequency table of **equivalence class** sizes.

False positive: An incorrect match between two data units or between a **data unit** and a **population unit**.

Formal Anonymisation: Any process which removes or masks **direct identifiers** on a dataset.

Formal identifier: Synonym of **direct identifier.**

Functional Anonymisation: A holistic approach to anonymisation which asserts that data can only be determined as anonymised or not in relation to its environment.

Guaranteed Anonymisation: A form of anonymisation where, given a set of assumptions, the risk of identification is zero. The most extreme form of this is where, given the environmental and data controls, an absolute risk of zero is claimed but this is widely thought of as a straw man.

Harmonisation: The process of recoding a variable on a dataset so that it more directly corresponds to an equivalent variable on another dataset.

Identifiable data: Data that contains indirect identifiers.

Identified data: Data that contains direct identifiers.

Impact management: A process which acknowledges that the risk of a disclosure from data that has been released or shared is not zero and therefore puts in place strategies to reduce the impact of such a disclosure should it happen.

Indirect identifiers: These can in principle include any piece of information (or combination of pieces of information). For example, consider a combination of information for a 'sixteen year old' and 'widowed'; whilst *age* and *marital status* are not immediately obvious identifiers, our implicit demographic knowledge tells us that this combination is rare. This means that such an individual could potentially be re-identified by, for example, someone spontaneously recognising that this record corresponded to someone they knew.

Informed consent: Basic ethical tenet of scientific research on human populations. Informed consent refers to a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including awareness of any risks involved, of uses and users of the data, and of alternatives to providing the data.

k-anonymity: A criterion sometimes used to ensure that there are at least k records within a dataset that have the same combination of indirect identifiers. Sometimes termed as using a threshold of k (usually 3 or 5 is used).

Key variable: A variable common to two (or more) datasets, which may therefore be used for linking records between them. More generally, in scenario analysis, the term is used to mean a variable likely to be accessible to the data intruder.

License agreement: A permit, issued under certain conditions which enables a researcher to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information.

Metadata level controls: Disclosure control methods that work by restricting the data rather than distorting it. Examples are sampling, variable deletion and aggregation/recoding.

Microaggregation: A form of disclosure control whereby data units are grouped based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates (perhaps group means or centroids) for those variables. The aggregates are released instead of the individual record values.

Microdata: A microdata set consists of a set of records containing information on individual data subjects. Each record may contain hundreds or even thousands of pieces of information.

Noise addition: The distortion of data through some random process.

Open data: Data released without any access restrictions, usually by publishing on the Internet.

Output statistical disclosure control: A process by which analytical outputs are manipulated so that they are non-personal. This is most relevant to data centres where access is controlled but the data are highly detailed and would be personal if released as open data.

Overimputation: Replacing real values in a micro-dataset with ones that have been generated through a statistical model.

Perturbation: Is a method for altering data in some way so as to control disclosure. Perturbative techniques include: **data swapping**, **noise addition, rounding** and **barnardisation.**

Penetration test: An approach to disclosure risk assessment where one attempts to re-identify individuals within a dataset using other (possibly publically available) information.

Personal data: Any information relating to an identified or identifiable data subject. An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous. Under Data Protection Act (1998), this term can only refer to living individuals. However, under other legislation the definition is extended to deceased individuals.

Personal information: A term used under Statistics and Registration Service Act (2007) – applying to that released by Office for National Statistics only – for information that either directly identifies an individual case or does so in conjunction with other information that is already in the public domain (published). Information for which identification requires privately held information does not constitute personal information. Personal information in this definition does include information about the dead as well as the living.

Population: the set of population units that a dataset is drawn from. The dataset could be a sample and so not all units within the population will necessarily be in the dataset.

Population unique: A record within a dataset which is unique within the population on a given set of key variables.

Population unit: An entity in the world. It is usually employed to mean the socio-physical analogue of a corresponding data unit although in any given dataset a given population unit may not have a corresponding data unit.

Privacy: A concept that is much discussed and debated and for which there is no unequivocal definition. It would be generally agreed that privacy applies to people whereas confidentiality applies to data. There is a definite relationship between confidentiality and privacy Breach of confidentiality can result in disclosure of data which harms the individual. This can be regarded as a violation of privacy because it is an intrusion into a person's self-determination (of the way his or her personal data are used). Informational (or data) privacy therefore can be understood to encompass an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

Pseudonymisation: A technique where direct identifiers are replaced with a fictitious name or code that uniquely identifies an individual; it is almost always used in conjunction with other **anonymisation** methods.

Quasi-identifier: Synonym of **indirect identifier**.

Record linkage: A process attempting to classify pairs of matches between different datasets.

Re-identification: The discovery of the identity of individual(s) in a dataset by using additional relevant information.

Remote access: On-line access to protected **microdata**.

Respondent: Originally used to refer to a person who responds to a survey. A respondent might provide data about just themselves but sometimes about others (as well) and data could have been generated without the data subjects' knowledge. So a respondent might not be a data subject and vice versa.

Response knowledge: The knowledge that a given **population unit** is included in a dataset. This could be through private knowledge, e.g. that a friend or work colleague has mentioned that s/he responded to a particular survey or it could be through simple knowledge that a particular population unit is a member of the population and the data is a full dataset for that population (e.g. a census).

Restricted access: A **data protection** measure that limits who has access to a particular dataset. Approved users can either have: (i) access to a whole range of raw

(protected) data and process it themselves or (ii) access to outputs, e.g. tables from the data.

Rounding: A method of statistical disclosure control where a figure is rounded off to a defined base; it is most commonly applied to tables of counts.

R-U (Risk-Utility) map: A graphical representation of the trade-off between **disclosure risk** and data utility.

Safe data: Data that has been protected by suitable Statistical Disclosure Control methods.

Safe setting: An environment such as a data lab whereby access to a disclosive dataset can be controlled.

Sample unit: A **data unit** in a dataset which is the sample of some **population**.

Scenario Analysis: A framework for establishing the key variables that might be used by a data intruder to re-identify data units.

Sample unique: A record within a dataset which is unique within that dataset on a given set of key variables.

Sampling: This refers to releasing only a proportion of the original data records on a microdata file. In the context of disclosure control, a data intruder could not be certain that any particular person was in the file.

Sampling fraction: The proportion of the population contained within a dataset. With simple random sampling, the sample fraction represents the proportion of population units that are selected in the sample. With more complex sampling methods, this is usually the ratio of the number of units in the sample to the number of units in the population from which the sample is selected. A low sampling fraction can provide some protection to a dataset where an intruder might not be able to infer that a **sample unique** is a **population unique**.

Scenario analysis: A framework for analysing plausible data intrusion attempts. This framework identifies (some) of the likely factors, conditions and mechanism for disclosure.

Secondary differentiation: A strategy adopted by a **data intruder**, to distinguish between multiple candidate matches between **data units** and **population units**. For multiple data units matched to a single population unit this involves identifying variables where the two records differ and then targeting resources on establishing the value of that variable for the population unit. For a single population unit matched against multiple population units this involves identifying which of the population units matches the data units on variables not included in the original match key.

Sensitive variables: Variables contained in a data record that belong to the private domain of data subjects who would not like them to be disclosed. There is no exact

definition given for what is a 'sensitive variable'. The Data Protection Act (DPA) lists twelve topics described as 'sensitive personal data' including: racial or ethnic origin, political opinions, religious beliefs, trade union membership, physical or mental health or condition, sexual life, and some aspects of criminal proceedings. However, there are other variables not in the DPA that might be deemed sensitive, such as those related to income, wealth, credit record and financial dealings. The context is important here; the distinction between sensitive and non-sensitive can depend on the circumstances. For example, one's religion might be considered as a sensitive variable in some countries and not so in others.

Special unique: A **sample unique** that has a high probability of being a **population unique**. This can be evaluated statistically and also through common sense knowledge. For example, intuitive knowledge of UK demographics will tell you that '16 year old widowers' are unusual. So if you have one in data for a particular geographical area then they may well be a population unique.

Statistical Disclosure Control (SDC): An umbrella term for the integrated processes of disclosure risk assessment, disclosure risk management and data utility.

Statistical disclosure: A statistical disclosure is a form of data **confidentiality** breach that occurs when, through statistical matching, an individual data subject is identified within an anonymised dataset and/or confidential information about them is revealed. A statistical disclosure may come about through: (i) the processes of **re-identification** and **attribution** (i.e. the revealing of new information) or (ii) the process of **attribution** alone.

Stream data: Data which is generated continuously either as an update or additively.

Subtraction attack: An attack carried out on aggregated data which works by removing completely known units from the data.

Suppression: A disclosure control process where parts of the data are made unavailable to the user. All metadata level controls could be viewed as a form of suppression but the term is more usually used to describe more targeted approaches like **cell suppression** and the removal of outliers and/or local suppression of particular values within microdata records.

Synthetic data: Data that have been generated from one or more population models, designed to be non-disclosive.

Tabular data: Aggregate information on entities presented in tables.

Target dataset: An anonymised dataset in which an intruder attempts to identify data subjects.

Target Variable: Within a scenario framework, information that an intruder would like to learn about a population unit or units.

Top coding: An SDC method used with interval-scale or ordinal variables where values above a certain threshold are aggregated together in order to mask a sparse

upper end of the distribution. Age and income are two variables that are typically treated in this way.

Appendix A: Standard Key Variables.

Standard keys are generated by organisations carrying out ongoing data environment analysis (scanning the data environment for new data sources). You should be aware that standard keys are generic and are set up primarily for use with licence-based dissemination of official statistics and will not be relevant to every data situation. If you are using a highly controlled access environment, or at the other end of the scale open data, or if you have data that is unusual in any way, this may not be the method to use.

However, the standard keys can be useful because if your data are not safe relative to these standards then in itself that indicates that you may have a problem, even before you consider non-standard keys.

The sets of keys presented here are subsets of those generated by the Data Environment Analysis Service at the University of Manchester using the methodology reported in Elliot et al (2011). They are focused on demographics and socio-economic variables. It should be stressed that these lists are time-dependent and are very much subject to change as the data environment changes. However, they will serve as a good starting point for considering your own data situation and its key variables.

Scenario Set A: Restricted access database linkage

Scenario A1.1: Restricted access database cross match (general)

This Scenario is based upon an analysis of the information commonly available in restricted access databases.

- Home address
- Age
- Sex
- Marital status
- Number of dependent children
- Distance of journey to work
- Number of earners
- Primary economic status
- SOCmajor (Standard Occupational code)

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

Scenario A1.2: Restricted access database cross match (general, extended)

This scenario is based upon an analysis of the information commonly available in restricted access databases, a slightly extended version of B1.1 with additional, less common variables. Typical variables are:

- Age
- Sex
- Marital status
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Number of earners
- Tenure
- Number of cars
- SOCmajor
- Primary economic status
- Income

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

Scenario A2.1: Restricted access database cross match (health)

This represents an attack from a restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness
- Type of primary long term illness (possibly match against multiple variables)

Attacker profile: Individual with access to restricted access dataset.

Scenario A2.2: Restricted access database cross match (health, extended)

This represents an attack from an extended restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness
- Type of primary long term illness (possibly match against multiple variables)
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Number of earners
- Tenure
- Number of cars
- SOCmajor
- Primary economic status

Attacker profile: Individual with access to restricted access dataset.

Scenario A3.1: Restricted database cross match (personnel)

This scenario is based on information commonly held in personnel databases. Typically this includes considerable detail on economic characteristics such as occupation, industry, economic status, basic physical characteristics (such as age, sex and ethnic group) and some information on personal circumstances (area of residence, long term illnesses, marital status and number of children).

- Home address
- Age
- Sex
- Marital status
- Primary economic position (filter)
- Occupation
- Industry
- Hours of work

- Migration in the last year
- Ethnic group
- long term illness
- Number of children.

Attacker Profile: Person working in personnel office of large organisation.

Scenario Set B: Publicly available information based attacks

Scenario B1.1: Commercial database cross match (common)

This scenario is based upon an analysis of the information commonly available in commercial databases. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Tenure
- Primary economic status
- Social grade
- Household composition

Attacker Profile: Person or organisation with sufficient resources to purchase lifestyle database type information.

Scenario B1.2: Commercial database cross match (superset, resource cost high)

This scenario is based upon an analysis of the information available in commercial databases. This is effectively a superset of available variables which could be exploited by a well-resourced attacker who links multiple data sources together.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Tenure
- Accommodation type
- Primary economic status

- Social grade
- Household composition
- Religion
- Number of rooms
- Income
- Transport to work
- Highest qualification
- Long term limiting illness
- Workplace

Attacker Profile: Person or organisation with sufficient resources to purchase multiple lifestyle databases.

Scenario B2: Local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variable age and ethnic group from the electoral register that could be used in a crude form are not included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Lowest floor in household
- Number of rooms
- Presence of bath
- Presence of central heating

Attacker Profile: Anyone.

Scenario B3: Extended local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variables (new voter/adult) and ethnic group that could be used in a crude form from the electoral register are included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Lowest floor in household
- Number of rooms
- Presence of bath
- Presence of central heating
- Ethnic group

- Age group (new voter/adult)

Attacker Profile: Anyone.

Scenario B4.1: Public information (low resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, and who is constrained in his/her use of resources.

- Home address
- Ethnic group (crude)
- Age
- Sex
- Qualifications
- Occupation
- Workplace

Scenario B4.2: Public information (high resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, without effective resource constraints.

- Home address
- Ethnic group (crude)
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Tenure
- Accommodation type

Scenario B4.3: Public information (high resources, opportunistic targeting attack)

This scenario imagines an intruder who is drawing on publicly available data sources, targeting a small number of individuals, who have visibility perhaps because of media coverage, without any resource constraints.

- Home address
- Ethnic group
- Age
- Sex
- Qualifications

- Occupation
- Workplace
- Tenure
- Accommodation type
- Marital status
- Country of birth
- Religion
- Nationality

Scenario B5.1: Online data sweep (low resources, opportunistic targeting attack)

This scenario envisages somebody trawling the net for available sources of information. The status of such information is questionable since much of it is deliberately self-published. For specific individuals the list of variables may be much longer than this. However, these will be commonly obtainable from online CVs and sites such as dating sites:

- Home address
- Ethnic group
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Marital status
- Dependents (y/n)
- Religion
- Income
- Language

Scenario B6.1: Worker using information about colleagues

This scenario is based upon a study of what people commonly know about people with whom they work. Typically this includes considerable detail on economic characteristics, basic physical characteristics and some very crude information about personal circumstances. Typical variables are:

- Age
- Sex
- Ethnic group
- Occupation
- Workplace
- Distance of journey to work

- Industry
- Hours
- Economic status
- Long Term illness
- Number of children

Attacker profile: Anyone working in a large organisation.

Scenario B6.2: Nosy neighbour

This scenario encompasses information that would be relatively easy to obtain by observation of one's neighbours. Obviously this does not entail either a standard match or fishing type attack. In effect one would be fishing for one's neighbours in the dataset. However if one found a match one could use information in the dataset to determine whether it is rare or not. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of Dependent children
- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multiethnic household
- Number of residents
- Number of rooms

Scenario B7.1: Combined public and visible sources

This is essentially the combination of nosy neighbour with publicly available information scenarios. This is quite a resource intensive attack because it involves hunting for information on a small group of people in public records. It is not likely to yield the information below on all neighbours.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children

- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multi-ethnic household
- Number of residents
- Number of rooms
- Qualifications
- Occupation
- Workplace
- Tenure
- Country of birth
- Religion
- Nationality

Scenario B7.2: Combined public, visible and commercial sources.

This is essentially the combination of nosy neighbour with publicly available information together with a superset of commercially available data. This implies a very well-resourced attacker who is carrying out a deep information gathering exercise on a small targeted population. Note the list of variables is more extensive than might be obtained on any restricted access database.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multi-ethnic household
- Number of residents
- Number of rooms
- Occupation
- Workplace

- Tenure
- Country of birth
- Religion
- Nationality
- Number of cars
- Number of dependent children
- Tenure
- Accommodation type
- Primary economic status
- Social grade
- Household composition
- Income
- Transport to work
- Highest qualification
- Long term limiting illness

Scenario Set C: Collusive attacks

Collusive attacks are ones where the data subjects collude in providing information about themselves. These do not intrinsically constitute a set against which a data controller is legally bound to protect. However, a successful collusive attack could still carry some risk, for example in terms of reputational damage.

Scenario C1.1: Demonstrative political attack: restricted set

The assumption underlying this scenario is that a political group, such as an anti-government group, acts in collusion with a data subject for the purpose of embarrassing the Government by undermining its data collection/release activities. Imagine that the data subject provides the group with copies of the information they gave to the interviewers. This scenario could happen in a census, which is a major public investment. Here the data collection process is familiar to everyone, and colluding respondents could be prepared in advance, and be guaranteed to be in the collected data (and also in the outputs with a relatively high probability). In principle, a larger number of variables could be used, but in the restricted variant, we have avoided those that are difficult to code (such as occupation), on the assumption that the political organisation will attempt to minimise divergence to prevent the demonstration backfiring. We have also avoided those that give information about other individuals apart from the colluding agent, on the assumption that the use of such variables would go against the underlying rationale for the attack.

- Home address
- Age
- Sex
- Education
- Marital status
- Primary economic status
- Ethnic group
- Religion
- Country of birth
- Migration in the last year
- Tenure
- Long term limiting illness
- Self-reported health
- Income

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.

Scenario C1.2: Demonstrative political attack: extended set

- Home address
- Age
- Sex
- Marital status
- Primary economic status
- Ethnic group
- Religion
- Country of birth
- Migration in the last year
- Long term limiting illness
- Self-reported health
- Income
- Number of rooms
- Tenure
- Housing type
- Number of residents
- Number of children

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.

Appendix B: Instructions for Calculating the Number of Uniques in a File.

These instructions assume that you have downloaded the appropriate data from the UKAN website (either Basetton.xlsx for Excel or Basetton.sav for SPSS) and have it open in the appropriate software. They also assume that you have a basic familiarity with the software package. The file is synthetic data but the data structure is that which might typically be found in a census, survey or administrative file.

In both cases we are using an eight variable key which represents information that somebody might plausibly know about a neighbour. You can play about with different variable combinations to see the impact on the number of uniques. Nothing should be read into the specific details of the results (the data is not real) – the exercises simply serve to demonstrate the technique which you can then use with your own data.

B.1 Instructions for Excel

1. Sort the file by the following columns (checking the 'my data has headers' box is checked): sex, age, ethnic, accomtype, tenure, marstatus, ncars, cenheat. For each column, sort from smallest to largest.
2. Enter the word 'ccount' into cell N1
3. Enter 1 in cell N2
4. Enter the following formula into cell N3
=IF(AND(A3=A2,B3=B2,C3=C2,D3=D2,E3=E2,F3=F2,J3=J2,M3=M2),N2+1,1)[96]
5. Fill down from N3 to N210745
6. Select and copy column N
7. Right click 'Paste' and pick the values option (ensuring the values are associated with the correct row as you carry out further sorting and calculations)
8. Repeat the sort you did at stage 1, but adding in ccount to the end of the list sorted from largest to smallest.
9. Enter the word 'csize' into cell O1

[96] This formula construction is based on the version of Excel available in the UK. We understand that in some countries that semi colons and used rather than commas in formula.

10. Enter the following formula into cell O2:
    =N2
11. Enter the following formula into cell O3:
    =IF (N3<N2,O2,N3)
12. Fill down from O3 to O210745
13. Switch to the output page tab
14. In cell B2 type the formula
    =COUNTIF (Barsetton!O:O,1)

B.2 Syntax for SPSS

```
SORT CASES BY sex(A) age(A) ethnic(A) accomtype(A) tenure(A) marstatus(A)
ncars(A) cenheat(A).

COMPUTE eccount=1.
IF (sex=lag(sex) & age=lag(age) & ethnic=lag(ethnic) & accomtype=lag(accomtype) &
tenure=lag(tenure) & marstatus=lag(marstatus) & ncars=lag(ncars) &
cenheat=lag(cenheat)) eccount=lag(eccount)+1.
EXECUTE.

SORT CASES BY sex(D) age(D) ethnic(D) accomtype(D) tenure(D) marstatus(D)
ncars(D) cenheat(D) eccount(D).
COMPUTE ecsize=eccount.
IF (eccount<lag(eccount)) ecsize=lag(ecsize).
EXECUTE.

COMPUTE unique=0.
VARIABLE LABELS unique 'Is the case unique?'.
VALUE LABELS unique 0 'No' 1 'Yes'.
IF (ecsize=1) unique=1.
EXECUTE.

FREQUENCIES VARIABLES=unique
 /ORDER=ANALYSIS.
```

Appendix C: A Description of the Data Intrusion Simulation (DIS) Method.

C.1 Introduction

The concept behind the DIS method derived from concerns expressed by Elliot (1996) regarding the need to examine statistical disclosure risk from the viewpoint of the data intruder (intruder-centrically) rather than from that of the data themselves (data-centrically). A rational intruder would be indifferent to questions such as, for example, whether a record was sample or population unique, because s/he will know such attributions of status are unreliable and more importantly because s/he will have more pragmatic concerns, such as whether her/his actual matches are correct. The DIS method simulates the intruder perspective by focusing on the probability of a unique match being correct. The basic assumption is that the intruder has some information about a population unit and uses that information to attempt to find the record for that individual in a microdata file (which is a sample of the relevant population). If there is only one record in the dataset which corresponds to the information that the intruder has that is called a unique match. If that record is the correct record for that population unit that is called a correct match. These basic elements form the headline statistic of a DIS analysis; the probability of a correct match given a unique match: pr(cm|um).

The basic principle of the DIS method is to remove records from the target microdata file and then re-sample them according to the original sampling fraction (the proportion of the population that are in the sample). This creates two files, a new, slightly truncated, target file and a file of the removed records which can then be matched against the target file. The method has two computational forms, the *special form*, where the sampling is actually done, and the *general form*, where the sampling is not actually performed, but its effect is derived using the equivalence class structure and sampling fraction.

C.2 The special method

The special DIS method uses a similar technique to Briggs (1992).

1. Set counters U and C to zero.
2. Take a sample microdata file (A) with sampling fraction S.
3. Remove a random record (R) from A, to make a new file (A').

4. Generate a random number (N) between 0 and 1. If N<=S then copy back R into A' with each record having a probability of being copied back equal to S.
5. The result of this procedure is that B will now represent an arbitrary population unit whose probability of being in A' is equal to the original sampling fraction.
6. Match fragment against A'. If R matches a single record in S' then add record 1 to U if the match is correct add 1 to C.
7. Iterate through stages ii-v until C/U stabilises.

C.3 The general method

A more general method can be derived from the above procedure. Imagine that the removed fragment (B) is just a single record. Clearly there are six possible outcomes depending on whether the record is resampled or not and whether it was a unique, in a pair, or in a larger equivalence class.

Table 1: Possible per record outcomes from the DIS general method

record is:	*Copied back*	*Not copied back*
sample unique	**correct unique match**	non-match
one of a sample pair	multiple match including correct	**false unique match**
one of a larger equivalence class	multiple match including correct	false multiple match

Given this, one can derive the estimated probability of a correct match given a unique match from:

$$\frac{U \times \Pi}{U \times \Pi + P \times (1 - \Pi)}$$

Where U is the number of sample uniques, P is the number of records in pairs and Π is the sampling fraction.

For full statistical proof of the above theory see Skinner and Elliot (2002). For a description of an empirical study that demonstrates that the method works see Elliot (2000). For an elaboration using the special method for post-perturbation disclosure risk assessment see Elliot (2001). For an extension which takes account of general misclassification errors see Elamir and Skinner (2006)

Appendix D: Instructions for Calculating the DIS Score.

These instructions assume that you have downloaded the appropriate data from the UKAN website (either Basetton sample.xlsx for Excel or Basetton sample.sav for SPSS) and have it open in the appropriate software. The file is synthetic data but the data structure is that which might typically be found in a census, survey or administrative file.

In both cases we are using an eight variable key which represents information that somebody might plausibly know about a neighbour. Nothing should be read into the details of the results (the data is not real) – the exercises simply serve to demonstrate the technique which you can then use with your own data.

In both cases we are using a file where the sampling faction is 10%.

D.1 Instructions for Excel

1. Sort the file by the following columns (checking the 'my data has headers' box is checked): sex, age, ethnic, accomtype, tenure, marstatus, ncars, cenheat. For each column, sort from smallest to largest.
2. Enter the word 'ccount' into cell N1
3. Enter 1 in cell N2
4. Enter the following formula into cell N3
   =IF(AND(A3=A2,B3=B2,C3=C2,D3=D2,E3=E2,F3=F2,J3=J2,M3=M2),N2+1,1)
5. Fill down from N3 to N210745
6. Select and copy column N
7. Right click 'Paste' and pick the values option (ensuring the values are associated with the correct row as you carry out further sorting and calculations)
8. Repeat the sort you did at stage 1, but adding ccount to the end of the list sorted from largest to smallest.
9. Enter the word 'csize' into cell O1
10. Enter the following formula into cell O2:
=N2
11. Enter the following formula into cell O3:
=IF(N3<N2,O2,N3)

12. Fill down from O3 to O210745
13. Switch to the output page tab
14. In cell B2 type the formula
=COUNTIF(BarsettonSample!O:O,1)
15. In cell B3 type the formula
=COUNTIF(BarsettonSample!O:O,2)
16. Enter the sample fraction 0.1 into cell B4
17. Enter the following formula into Cell B5
=B2*B4/(B2*B4+B3*(1-B4))

D.2 Instructions for SPSS

The syntax to use is shown below. When you have run it you will have a frequency table which will give you counts for the number of unique records and the number which are members of identical pairs. You simply need to insert those numbers into the standard DIS formula:

$$\Pr(\text{cm} \mid \text{um}) = \frac{U \times \Pi}{U \times \Pi + P \times (1 - \Pi)}$$

Where U is the number of sample uniques, P is the number of records in pairs and Π is the sampling fraction, in this case 0.1.

SPSS syntax

```
SORT CASES BY sex(A) age(A) ethnic(A) accomtype(A) tenure(A) marstatus(A)
ncars(A) cenheat(A).

COMPUTE eccount=1.
IF (sex=lag(sex) & age=lag(age) & ethnic=lag(ethnic) & accomtype=lag(accomtype) &
tenure=lag(tenure) & marstatus=lag(marstatus) & ncars=lag(ncars) &
cenheat=lag(cenheat)) eccount=lag(eccount)+1.
EXECUTE.

SORT CASES BY sex(D) age(D) ethnic(D) accomtype(D) tenure(D) marstatus(D)
ncars(D) cenheat(D) eccount(D).

COMPUTE ecsize=eccount.
IF (eccount<lag(eccount)) ecsize=lag(ecsize).
EXECUTE.
```

```
COMPUTE uniquepair=0.
VARIABLE LABELS uniquepair 'Is the case unique or a one of a pair ?'.
VALUE LABELS uniquepair 0 'Not unique or pair' 1 'Unique' 2 'One of a pair'.
IF (ecsize=1) uniquepair=1.
IF (ecsize=2) uniquepair=2.
EXECUTE.

FREQUENCIES VARIABLES=uniquepair.
 /ORDER=ANALYSIS.
```

Appendix E: Data Features Template.

Feature type	Question	Answer/Actions
Data Subjects	Who are they?	
	What is their relationship with the data?	
Data type	Microdata, Aggregates or something else?	
Variable Types	What common indirect identifiers do you have?	
	What sensitive variables do you have?	
Data properties	Is the data accurate?	
	How old is the data?	
	Is it Hierarchical or flat?	
	Is it Longitudinal or Cross-sectional?	
	Population or sample (what fraction)	
Anything else of note?		

"This authoritative and accessible decision-making framework will help the information professional to anonymise personal data effectively. The framework forms an excellent companion piece to the ICO's code of practice."

Elizabeth Denham – UK information Commissioner

"It is my belief that this book will come to be seen as gold-standards in the field: it is fundamentally rational, scientifically and technically rigorous, easily understandable and framed in a way that makes them useable. I intend for it to become mandatory reading across my research group."

Paul Burton - Professor of Infrastructural Epidemiology at the University of Bristol

"The Anonymisation Decision Making Framework provides a practical and useful structure to assist data custodians in the process of anonymising a dataset before release or sharing. The lack of a practical guide providing a broadly applicable "how to" manual has been a noticeable gap until the appearance of this book. In addition, the book integrates (UK-specific) legal and technical perspectives to provide a comprehensive and readable treatment covering both the data and the data environment or context."

Christine O'Keefe – Senior Principal Research Scientist, Data61, CSIRO,
and Adjunct Professor, School of Mathematical Sciences, University of Adelaide.

"Extremely useful to data controllers who need to safely releaselshare sensitive personal data"

Josep Domingo-Ferrer
Professor of Computer Science at Universitat Rovira I Virgili and
UNESCO chair of Data Privacy

"The holistic approach proposed in this book is intuitive and sensible. I fully agree with the authors that looking at the anonymisation problem only from the data perspective falls short of important aspects of the problem. The authors illustrate clearly that considering the environment is at least as important as the data themselves."

Jörg Drechsler - Institut für Arbeitsmarkt und Berufsforschung, Germany