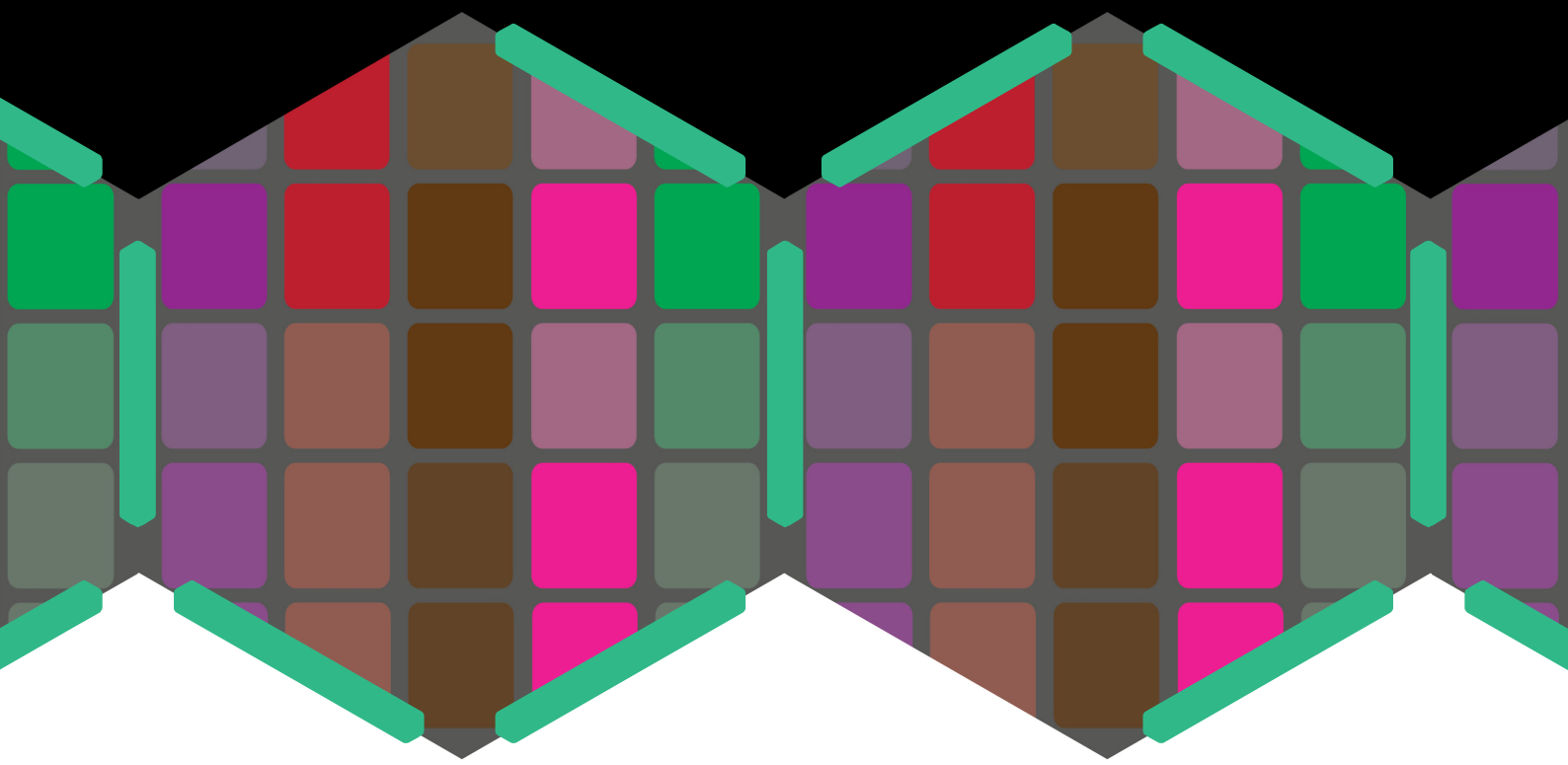
The De-Identification Decision-Making Framework

Christine M O'Keefe, Stephanie Otorepec,
Mark Elliot, Elaine Mackey and Kieron O'Hara

18 September 2017



Australian Government
Office of the Australian Information Commissioner

Citation

CM O'Keefe, S Otorepec, M Elliot, E Mackey, and K O'Hara (2017) The De-Identification Decision-Making Framework. CSIRO Reports EP173122 and EP175702.

This work is an adaptation to the Australian context of the publication: M Elliot, E Mackey, K O'Hara, and C Tudor. The Anonymisation Decision-Making Framework. UKAN Publications, UK, 2016. http://ukanon.net/ukan-resources/ukan-decision-making-framework/

Author affiliations

Christine M O'Keefe, CSIRO, Australia
Stephanie Otorepec, Office of the Australian information Commissioner, Australia
Mark Elliot, University of Manchester, UK
Elaine Mackey, University of Manchester, UK
Kieron O'Hara, University of Southampton, UK

Licensing and Copyright

Important disclaimer

Foreword

The quantity of data produced globally is increasing exponentially. At the same time, technological advances have highlighted the potential of this new data to positively transform all spheres of life. But the value of data can only be realised if it is able to be shared with those who have the necessary expertise to analyse and use it. This has both legal and ethical implications when the data contains personal information. How can data be shared consistently with both the standards set out in privacy (or personal data protection) legislation, as well as the broader standards and expectations of the community?

De-identification offers one potential solution to this problem. When done properly, it allows data to be shared or released with a wider audience in ways that protect individual privacy, and which may not otherwise be permitted under privacy legislation. De-identification is therefore a topic of huge interest to organisations around the world, and across both the public and private sectors.

While a growing body of de-identification literature has been appearing for some decades now, de-identification practice remains fairly inconsistent, and at times, unfortunately lacking in rigour. The use or release of poorly de-identified data gives rise to a range of serious privacy and other risks for affected data subjects. It also exposes an organisation to significant legal and reputational risks and can undermine trust, impeding the progress of future data-based projects. This is regrettable in cases where there may be a strong public interest justification for sharing data.

My Office has long recognised the importance of making data more available, subject to appropriate safeguards. We have assisted many organisations in navigating the line between access to valuable government-held information and the protection of personal privacy. As this resource goes on to demonstrate, this is no longer straightforward or two-dimensional. A balance must be struck at the intersection of law, policy, and technology.

I am therefore delighted to present this authoritative, comprehensive and accessible resource, which is the product of a close collaboration between my Office and the CSIRO Data61, with input from the Australian Bureau of Statistics and the Australian Institute of Health and Welfare. It has been adapted from the original UK version to reflect the legal framework of the Australian *Privacy Act 1988* (Cth), while remaining true to the original UK work which aims to give a fresh and integrated perspective on all aspects of de-identification practice.

This book is not intended to eliminate the need to 'call in the experts'. Indeed, expert advice - particularly on many of the technical aspects of de-identification - may be crucial. However, this book will empower your organisation to understand what is involved in a de-identification process, and how to begin identifying, evaluating and balancing the relevant risks effectively and realistically. In doing so, the book acknowledges that it is often difficult to draw a sharp distinction between personal information and de-identified data. De-identification is an exercise in risk management, rather than an exact science.

The framework set out in this book is fully consistent with my Office's guidance on de-identification. However, like the original UK version, it goes beyond the strict legal obligations and sets out a framework which flags a wide range of issues that should be considered during a de-identification process. For example, whether your organisation has a social licence for its data-based activities, and how to build trust through transparency.

The fundamental premise underpinning the framework is that re-identification risk must be assessed contextually. In line with this, this book expounds functional de-identification – a method which encourages data custodians to consider not only the data itself, but also the environment the data will be released into. Only through a full consideration of both factors can a data custodian determine which techniques and controls are necessary and appropriate to produce data that is both safe (meets legal obligations) and useful (sufficiently rich or detailed to meet the use case).

In all cases, for data to be considered 'de-identified', the risk of re-identification in the relevant release context must be very low. This requires custodians to balance data utility and the level of data modification required to achieve de-identification, while considering the most appropriate data environment which can facilitate this. There is sometimes an unavoidable trade-off here. Sharing data with a wider audience - or indeed, completely openly - may necessitate significant alterations to the data itself, which may lower data utility considerably. Importantly, the book notes that open data environments are really only appropriate for data that is either not derived from personal information, or data that through an extremely robust de-identification process that ensures with a very high degree of confidence that no individuals are reasonably identifiable.

Data custodians should therefore be aware that the application of environmental controls, which go to the 'who', 'what', 'where', and 'how' of accessing data, may be more effective at reducing the risk of re-identification than modifying the data itself, and with a lower impact on the utility of the data.

This book represents one of many potential ways to approach de-identification consistently with the Privacy Act. Our hope is that this book will generate further discussion, research and interest in this field, and we look forward to discussions with stakeholders about the future of de-identification best practice.

Timothy Pilgrim PSM

Australian Information Commissioner

Australian Privacy Commissioner

Preface

The need for well-thought-out de-identification has never been more acute.

On the one hand, there is an increasing recognition of the value of sharing and releasing data. As just one example, the Productivity Commission wrote:

> *Extraordinary growth in data generation and usability have enabled a kaleidoscope of new business models, products and insights to emerge. Individuals, businesses, governments and the broader community have all benefited from these changes (Productivity Commission 2017).*

On the other hand, there is a growing body of examples of data releases that have led to privacy concerns, including the Netflix, AOL and New York taxi cases (Arrington 2006, Atokar 2014, CNN Money 2005). Australian examples include the release and subsequent retraction of `a linkable de-identified 10% sample of Medicare Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS)' and the results of the most recent Australian Public Service Commission employee census (itnews 2016a, 2016b). While it can be debated whether or not the de-identification methods used were adequate in these examples, they clearly underline how important it is to ensure that de-identification is both carried out properly, and seen to be carried out properly.

The De-Identification Decision-Making Framework is an adaptation to the Australian context of the UK resource The Anonymisation Decision-Making Framework, and is the result of a close collaboration of CSIRO with the Office of the Australian Information Commissioner and the Australian Bureau of Statistics. The adaptation has required revisions due to differences in the legal frameworks, the use of Australian examples and terminology, and minimal additional changes. One major change from the UK resource is the substitution of the term 'de-identification' for 'anonymisation' in the original.

We have released this resource as a freely available open source book as we feel that we have an important message that we want to ensure is disseminated as widely as possible. We hope that you find the book of value.

Contents

Figures

Tables

Executive Summary

Aims, intended audience, and how to use this book

Aims

This book has been developed as a practical guide to de-identification, focussing on operational advice. It is intended to provide a pragmatic understanding of the de-identification process, and an idea about how to utilise it to advance business or organisational goals.

The book presents a generic approach to the process of de-identification which will help you to identify and address the key factors relevant to your particular data sharing or release situation.

Intended audience

The book is intended for organisations who have data about individuals that requires de-identification including Australian government agencies, not-for-profit and private sector organisations. There are a number of reasons why a data custodian might want to de-identify personal information, including: to advance business or organisational goals, to enable the sharing or releasing of data to realise social, environmental or economic value, and to fulfil legal obligations including those imposed by the *Australian Privacy Act 1988* (Cth) (Privacy Act), as well as any other applicable legislation.

How to use this book

The De-Identification Decision-Making Framework has been designed as a standalone guide to the de-identification process, providing a principled method for evaluating any data situation and then selecting appropriate environment-based and data-based controls to manage risk in that situation. Therefore, it can be used by itself without reference to the Five Safes or any other framework.

We recognise that the Five Safes framework has been adopted by a growing number of Australian government agencies as a model for managing disclosure risk. If you wish to use the Five Safes, for example in your communications with senior managers or the public, then after completing the De-Identification Decision-Making Framework you will be able to map your selections of environment-based and data-based controls onto the Five Safes, and verify that each of the Safes has been considered.

De-identification is not an exact science and, even using the De-Identification Decision-Making Framework (DDF) at this level, you will not be able to avoid the need for complex judgement calls. You are still likely to need expert advice on some parts of the de-identification process, particularly with the more technical risk analysis and control activities.

Key messages

Section 1

In this book, we use the term ‘de-identified’ to describe data that is not (or is no longer) about an identified or reasonably identifiable individual, and does not reveal personal information about such an individual. The term is used with the understanding that ‘de-identified’ is used in the same spirit as 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 it will have a negligible risk of collapsing.

De-identification is a process of risk management, but also a decision-making process designed to help answer the question: should we share or release this data and if so in what form and setting?

The framework is underpinned by the belief that you must look at both the data and the data environment to ascertain realistic measures of risk. Attention is thus shifted away from the traditional question ‘how risky is the data for release?’ towards the more critical question ‘how might a disclosure occur?’ The approach that we take here also includes the actions of other key agents, other data within the environment, and governance processes. The basic premise is that you cannot guard against the threat to de-identification unless you have a clear idea of what it is you are guarding against - and this requires considering both data and its environment.

The De-Identification Decision-Making Framework is based on five key principles:

1. It is impossible to decide whether data is safe to share/release by looking at the data alone.
2. But it is still essential to look at the data.
3. De-identification is a process to produce safe data but it only makes sense if safe *useful* data is produced.
4. Zero risk is not a realistic possibility in producing useful data.
5. The measures put in place to manage risk should be proportional to the risk and its likely impact.

Section 2

Under the Privacy Act, de-identification is a process that renders personal information which would otherwise be subject to the Privacy Act, into a form that is not identifiable. This releases it from such restrictions and allows it to be shared or disseminated, or put to uses which may otherwise not be permitted. For the purposes of the Privacy Act, information is de-identified if the risk of re-identification occurring is very low (having regard to the relevant release context).

In addition to the legal considerations, this book examines the relevant ethical concerns. There are two main reasons why ethical considerations are important, namely:

1. Data subjects may not want data about them being re-used in general, by specific third parties, or for particular purposes.
2. After de-identification, risk is still generally not zero.

Section 3

De-identification is a complex topic with many different components, and simply considering one aspect in isolation (for example, data confidentiality) could lead to difficulties, and a non-usable solution. There are two main approaches designed to assist custodians to navigate the complexities associated with de-identification. The first is Functional De-Identification, the basis of The De-Identification Decision-Making Framework. The second is the Five Safes framework, currently gaining popularity amongst Australian custodians.

Functional de-identification considers the whole of the data situation, i.e. both the data and the data environment. The objective is to ensure that de-identified data remains de-identified once it is shared or released within or into a new data environment (such as another agency, or a publicly accessible data portal). If de-identification is to be a useful tool for risk management, one has to specify its circumstances. Under Functional de-identification the question: 'is this data personal' requires an answer to the additional question: 'in what context?' or more specifically 'in what data environment?' A data environment is made up of four components: data, agency, governance processes and infrastructure.

1. **Other data**: What (other) data exists in the data environment? How does it overlap with or connect to the data in question?
2. **Agency**: Who is capable of acting on the data and in the data environment?
3. **Governance processes**: How are users' relationships with the data managed?
4. **Infrastructure**: How do infrastructure and wider social and economic structures shape the data environment?

A data situation captures the relationship between data and its environment, and functional de-identification is a process which controls disclosure risk by considering the totality of a data situation.

Consistent with the functional de-identification approach, the Five Safes is a framework for organising thinking about data access. The basic premise of the framework is that data access can be seen as a set of five 'risk dimensions': safe projects, safe people, safe data, safe settings, safe outputs. Each dimension provokes a question about access:

Safe projects: Is this use of the data appropriate?

Safe people: Can the researchers be trusted to use it in an appropriate manner?

Safe data: Is there a disclosure risk in the data itself?

Safe settings: Does the access facility limit unauthorised use?

Safe outputs: Are the statistical results non-disclosive?

These dimensions embody a range of values: 'safety' is a measure, not a state. For example, 'safe data' is the dimension under which the safety of the data is being assessed; it does not mean that the data is absolutely non-disclosive. Nor does it necessarily specify how the dimensions should be calibrated. 'Safe data' could be classified using a statistical model of re-identification risk, or a much more subjective scale, from 'very low' to 'very high'. The point is that the user has some idea of what is 'more safe data' and 'less safe data'.

While it may seem at first that there is a wealth of ways that one can share or release data outside organisational boundaries, in a manner that renders it de-identified, in fact there are four main options:

1. Open access: making data freely and publicly available, for example, on a web page

2. Delivered access: requested data is delivered to approved users under specified conditions

3. On-site safe settings: on approval, data is accessed in a secure, controlled location

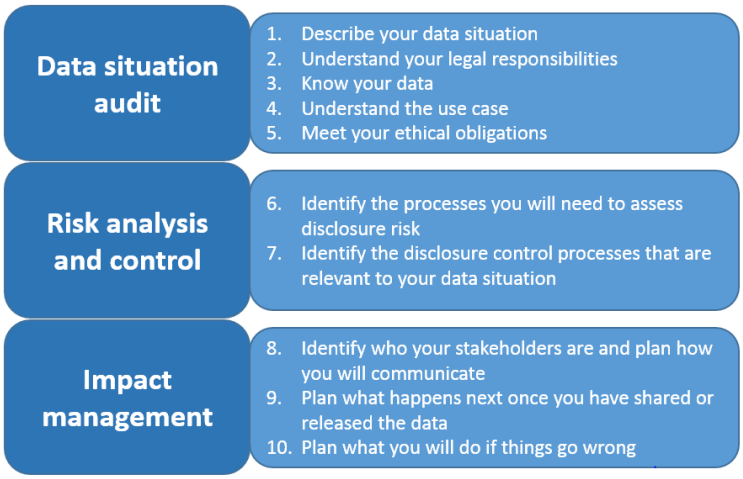
4. Secure virtual access: on approval, data is accessed via a secure link

Each of these options can be fine-tuned by implementing different governance processes including approvals systems, and infrastructure including researcher agreements and security measures such as secure storage and transfers, and audit trails.

Open data environments are really only appropriate to data that is either not personal in the first place or have been through an extremely robust data-focussed de-identification process that ensures with a very high degree of confidence that no individual could be re-identified and no disclosure could happen under any circumstances.

Summary of the De-Identification Decision-Making Framework

The De-Identification Decision-Making Framework (DDF) comprises ten components, or activities, from 1. Describe your data situation to 10. Plan what you will do if things go wrong. These ten components are grouped into three core de-identification activities, as shown in the following diagram:



A data situation audit (Components 1-5) 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 it and the issues thereby raised. A well-conducted data situation audit is the basis for the next core activity.

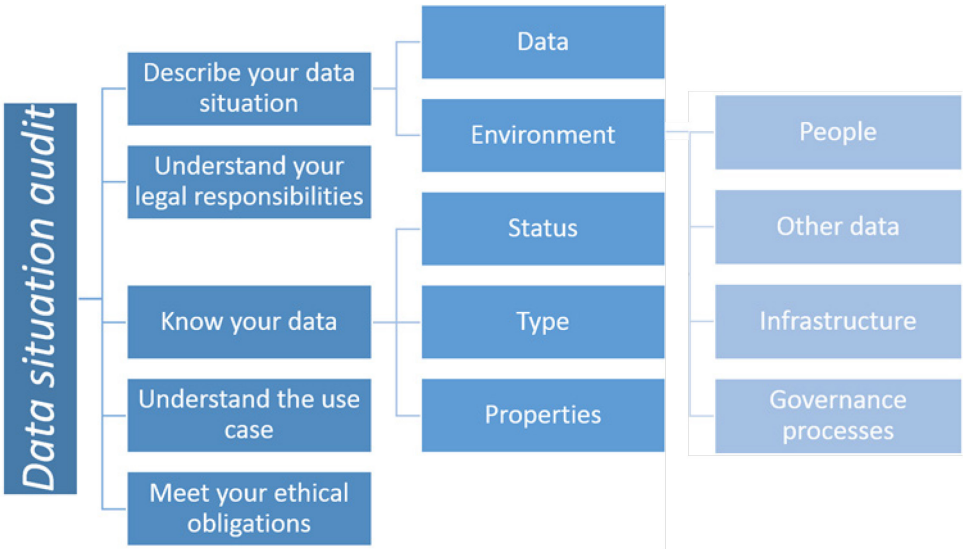
Risk analysis and control (Components 6-7) requires you to consider the technical processes that you will need to use in order to both assess and manage the disclosure risk associated with your data situation.

Impact management (Components 8-10) requires you to 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.

The data situation audit

The data situation audit is essentially a framing tool for understanding the relationship between your data and its environment, and therefore to help scope the de-identification process appropriately for you to share or release your data safely. It will also help you to clarify the goals of the process and will enable the more technical aspects of the de-identification process to be planned and conducted more rigorously.

The next figure shows a diagrammatic representation of the data situation audit, the first part of the De-Identification Decision-Making Framework.



Component 1: Describe your data situation

Your data situation comprises the data, as well as the other data, people, infrastructure and governance that make up its environment. There is often more than one data situation involved, such as if the data is being transferred from one organisation to another, or being released as open data.

Component 2: Understand your legal responsibilities

With regard to the Privacy Act, the key questions are:

1. is the data personal information or de-identified data, and

2. if it is de-identified data, what controls need to be in place to maintain this status?

Component 3: Know your data

Conduct a high-level examination of your data, focussing on the data type, features, and properties. This involves the data subjects, variables, quality, and age.

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 release and what is the most appropriate means by which to do this.

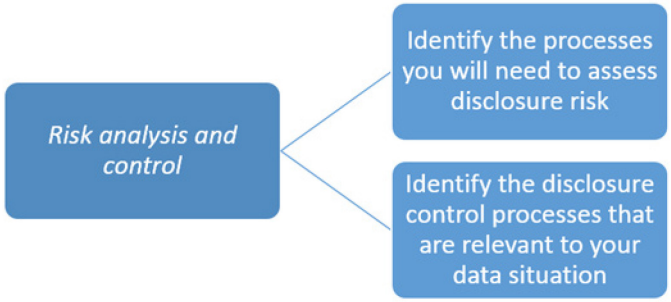
Component 5: Meet your ethical obligations

Considerations here include: consent, transparency, stakeholder engagement, and governance.

Disclosure risk assessment and control

Risk assessment and control should usually be an iterative, not linear, process. They will be constrained by the use case and the resources available.

The next figure shows a diagrammatic representation of disclosure risk assessment and control, the second part of the De-Identification Decision-Making Framework.



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

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 procedure 2 by simulating attacks using ‘friendly’ intruders.

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

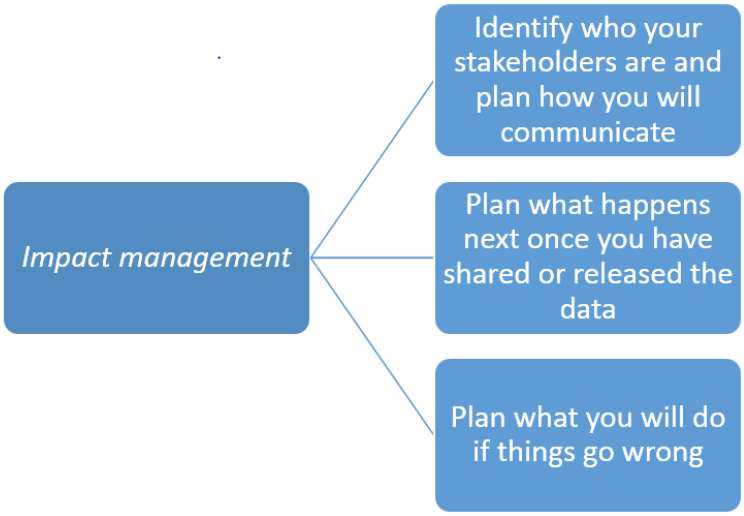
Disclosure control processes essentially attend to either or both of the two elements of your data situation: the data and its environment. If your risk analysis in Component 6 suggests that you need stronger controls then you have two (non- exclusive) choices:

1. Reconfigure the data environment
2. Modify the data, including possibly reducing the amount of data under consideration

Impact Management

Impact management puts in place a plan for reducing the impact of an unintended disclosure should it happen.

The next figure shows a diagrammatic representation of impact management, the third part of the De-Identification Decision-Making Framework.



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 trusted. You will be better placed to manage the impact of a disclosure if you and your stakeholders have developed a good working relationship.

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

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

1. Keeping a register of all the data you have shared or released, including a description of the associated data environment(s).
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.

3. Be aware of changes in the data environment and how these may impact on your data. This means:
   - keeping abreast of developments in new technologies that may affect your data situation by, for example, reading technology journals/blogs, watching relevant podcasts and/or attending relevant events,
   - monitoring changes in the law or guidance on data sharing and dissemination by engaging with relevant organisations such as the ABS, AIHW and/or OAIC, and
   - 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.

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

Sometimes, even when you follow best practice, things can go wrong. It is essential to put in place mechanisms that can help you deal with a disclosure in the rare event that one were to occur. Such measures include having: a robust audit trail, a crisis management policy, and adequately trained staff.

1 Introduction

In this section, we introduce our presentation of the De-Identification Decision-Making Framework (DDF) commencing with a description of the aims, and intended audience for this book with some guidance on how to use it. After a brief discussion of the related notions of de-identification, risk and sensitivity, we present the principles upon which the DDF is founded.

In this book, the term de-identified is used to describe data that is not (or is no longer) about an identified or reasonably identifiable individual, and does not reveal personal information about such an individual. For a deeper discussion of terminology, see Appendix A.

1.1 Aims, intended audience, and how to use this book

1.1.1 Aims

This book has been developed as a practical guide to de-identification, focussing on operational advice. It is intended to provide a pragmatic understanding of the de-identification process, and ideas about how to utilise it to advance business or organisational goals.

De-identification 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 with many uncertainties. It requires thinking about a range of heterogeneous issues from ethical and legal obligations to technical data questions, and integrating the different perspectives on the topic of de-identification into a single comprehensible framework is what this book is all about.

The book presents a generic approach to the process of de-identification which will help you to identify and address the key factors relevant to your particular data sharing or release situation. The biggest demand for this sort of guidance comes from people and organisations dealing with data that contains personal information and so that is the focus of our exposition. For example, we have not addressed the specialist topic of data about businesses, since business data has different technical properties and different legislation can apply to it. That said, the framework we present can be applied, perhaps with minor modifications to the detail, to just about any data where privacy or confidentiality is an issue but sharing is valuable.

1.1.2 Intended audience

The book is intended for organisations who have data about individuals that requires de-identification. There are a number of reasons why a data custodian might want to de-identify personal information, including: to advance business or organisational goals, to enable the sharing

or releasing of data to realise social, environmental or economic value, and to fulfil legal obligations including those imposed by the Australian *Privacy Act 1988* (Cth) (Privacy Act).[1]

The range of such data custodians with data to de-identify includes Australian government agencies as well as private sector organisations. For example, the Australian Government Department of Prime Minister & Cabinet's *Process for Publishing Sensitive Unit Record Level Public Data as Open Data*[2] requires the data custodian to determine a methodology to confidentialise (that is, de-identify) a dataset. The framework in this book can help in this endeavour.

The book may also be of interest to a de-identification specialist who would appreciate a fresh, integrated perspective on the topic.

1.1.3 How to use this book

The De-Identification Decision-Making Framework has been designed as a standalone guide to the de-identification process, providing a principled method for evaluating any data situation and then selecting appropriate environment-based and data-based controls to manage risk in that situation. Therefore, it can be used by itself without reference to the Five Safes or any other 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. Other more technical material appears in the Appendices.

We recognise that the Five Safes framework has been adopted by a growing number of Australian government agencies as a model for managing disclosure risk. Although it was not covered in the original UK version of this book, we have included an introduction. If you wish to use the Five Safes, for example in your communications with senior managers or the public, then after completing the De-Identification Decision-Making Framework you will be able to map your selections of environment-based and data-based controls onto the Five Safes, and verify that each of the Safes has been considered. We remark that the Five Safes may not explicitly cover all aspects of disclosure risk management that are part of the De-Identification Decision-Making Framework, such as deciding on whether or not to release data, what to do after it is released and what to do if something goes wrong.

De-identification is not an exact science and, even using the De-Identification Decision-Making Framework (DDF) at this level, you will not be able to avoid the need for complex judgement calls about when data is sufficiently de-identified given your data situation (that is, your data and its interactions with factors such as other data, people, governance, and infrastructure). The DDF will help you in making sound decisions based on best practice, but it is not a step-by-step algorithm; it is an approach whose value depends on the extent of the knowledge and skills you bring to it. You are still likely to need expert advice on some parts of the de-identification process, particularly with the more technical risk analysis and control activities.

---

[1] We use the most current compilation registered on 25 Oct 2016. See https://www.legislation.gov.au/Series/C2004A03712

[2] Available at: https://blog.data.gov.au/sites/g/files/net626/f/process_for_publishing_open_data_dec16.pdf

1.2 De-identification, risk and sensitivity

A common error when thinking about de-identification is to focus on a fixed end state of the data. This is a problem because it leads to much confusion about what it means to produce 'de-identified data'. First, it focuses exclusively on the properties of the data alone, whereas in reality whether data is de-identified or not is a function of both the data itself and its interaction with the surrounding data environment (collectively called the data situation). Second, it leads one into some unhelpful discussions about the relationship between de-identification and its companion concept risk, with some commentators mistakenly (or at least optimistically) assuming that 'de-identified' means that there is zero risk of an individual being re-identified within a dataset. Third, viewing de-identification as an end state means that one might assume that after a de-identification process, one's work is done which in turn promotes a counterproductive assumption that one can 'release-and-forget'.

In some ways, it would be better to drop the (adjectival) form 'de-identified' altogether and perhaps talk instead of 'data that has been through a de-identification 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 'de-identified' with the understanding that it is used in the same spirit as 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 it 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 (Tversky and Kahneman 1974). One aspect of this is the tendency to confuse the likelihood of an event with its impact. To complicate matters further, where risks are dependent on human action, these biases themselves factor into the risk profile. For example, if we can convince a data intruder that the likelihood of a disclosure attempt succeeding is negligible then they are less likely to expend the necessary effort 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 privacy breach. However, as we will see, sensitivity is a larger concept than this and in the context of data it also encompasses how the data was collected and what reasonable expectations a data subject might hold about what will happen to data about them.

De-identification, 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 with many uncertainties. It does require thinking about a range of heterogeneous issues from ethical and legal obligations to technical data questions, and bringing all these disparate elements into a single comprehensible framework is what this book is all about.

1.3 The principles behind the DDF

The DDF incorporates two frames of action: one technical, the other contextual. The technical element of the framework will enable you to think about both how to quantify disclosure risk and how to manage it. The contextual element will enable you to think about and address those factors that affect disclosure risk. These include the particulars of your data situation such as the data flows, 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 the belief that you must look at both the data and the data environment to ascertain realistic measures of risk. Attention should be shifted away from the traditional question 'how risky are the data for release?' towards the more critical question 'how might a disclosure occur?' This approach to risk assessment and management considers not just of the data to be released but also the actions of 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 de-identification unless you have a clear idea of what it is you are guarding against - and this requires considering both data and its environment.

What this means for you is that your assessment and management of disclosure risk should include reference to all the components of the DDF, including your data, other external data sources, legitimate data use and potential misuse, governance practices, and your legal, ethical and ongoing responsibilities. The DDF 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
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 Section 4. What we will do is make explicit the five principles upon which the DDF is founded, as follow.

1. **You cannot decide whether data is safe to share/release 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 and 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. **De-identification is a process to produce safe data but it only makes sense if what you are producing is safe <u>useful</u> data:** You may wonder why we talk about the need to balance data utility with data safety in the de-identification process. It is easy after all to think about de-identification only in terms of producing safe data but if you do that you may well be taking a risk for no benefit. Remember, de-identification 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 does not represent whatever they are meant to represent. There are two possible outcomes that arise from low utility and neither are happy ones:
     - the data is of little or no use to its potential users and you will have wasted your time and resources on them, or
     - the data could lead to misleading conclusions which might have significant consequences if, for example, the data is used to influence policy or to make decisions.
   - 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. De-identification 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 privacy 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 is 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 facility.
   - If the data has minimal detail and is 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.

1.4 Structure of this book

In this section we have introduced some of the core concepts relevant to our approach to de-identification. We have also provided a top level overview of the De-Identification Decision-Making Framework, explaining both the thinking behind it and the principles on which it is founded. The DDF is a generic approach to the process of de-identification which will help you to identify and address the key factors relevant to your particular data sharing or release situation.

In the next section we discuss in depth how de-identification relates to the Privacy Act, and the importance of ethics in de-identification.

Section 3 brings together a small number of fundamental ideas and concepts required to understand and implement the DDF.

In Section 4 we present the DDF, working through each component in detail. The approach 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. In most 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 in this book we have made liberal use of footnotes. Our intention is that the book can be read and the framework understood without paying too much attention to the footnotes at all – they are there for those who may want references or more detail.

In order to keep the book focussed on the DDF process, we have moved some of the more technical material to the appendices. This includes a further discussion of terminology, as well as more comprehensive introductions to data-based methods for disclosure risk measurement and reduction. The appendices also contain lists of standard key variables appropriate to a range of attack scenarios, as well as other information helpful in implementation. Where it is relevant in the text, we have let the reader know where to find more detail.

2 De-Identification, privacy and ethics

In this section we will discuss in depth how de-identification relates to the Privacy Act, and the importance of ethics in de-identification.

2.1 De-identification and the Privacy Act

Under the Privacy Act, as we shall see in this section, properly carried out de-identification renders personal information which would otherwise be subject to the Privacy Act, as not identifiable, thereby releasing it from such restrictions and allowing it to be shared or disseminated, or put to uses which may otherwise not be permitted. It therefore allows data to be shared legally (and more ethically), and can help facilitate the realisation of data's huge social, environmental and economic value.

2.1.1 The origins of data protection laws, and identifiability as the basis of regulation

Privacy is recognised as a fundamental human right. Around the world, many jurisdictions have enacted privacy or data protection laws, in large part to meet their obligations under relevant international human rights treaties.[3] The meaning of 'privacy' as a broader concept is highly contested.[4] However, most data protection or privacy laws (including Australia's Privacy Act) regulate information privacy only - they do not define (or purport to protect) all aspects of the broader notion of 'privacy'.

As outlined in the introduction, most data protection laws are centred on the notion of identifiability. They do not place restrictions on the handling of any data about individuals - only identifiable data. By way of example, such laws may impose:

- Restrictions on use: identifiable data should generally only be used for the same purpose for which it was obtained
- Restrictions on sharing/releasing: identifiable data should generally only be given to others where this is consistent with the purpose for which it was obtained
- Obligations to secure data – identifiable data should be protected from theft, loss, or unauthorised access, and
- Obligations to allow individuals to access and correct their data.

[3] See Article 17 of the International Covenant on Civil and Political Rights (ICCPR), which reflects Article 12 of the 1948 Universal Declaration of Human Rights.

[4] It is a somewhat amorphous concept that implicates psychological notions like identity and autonomy, and depends on a locus of control which is contextualised by cultural norms. This broader conception of privacy may be important for entities to consider when deciding whether a particular use of data is, overall, ethical and acceptable to the public, or to the data subjects themselves.

This is just a sample of the types of obligations which apply to personal information/identifiable data across a number of jurisdictions.[5]

2.1.2 Privacy law in Australia

In Australia, the Privacy Act regulates the protection of personal information (or data which contains personal information). It applies to most Australian Government agencies, all private sector and not-for-profit organisations with an annual turnover of more than $3 million, all private health service providers, and some small businesses (sometimes called 'APP entities'). It contains thirteen 'Australian Privacy Principles' (APPs) which regulate all aspects of information-handling throughout the information cycle. That is, from the point when the APP entity obtains the information, through to its use, storage and finally its disposal.[6] For the purposes of this section, we will assume that all data custodians are APP entities.

The obligations contained in the Privacy Act are triggered when a data custodian 'collects'[7] or 'holds'[8] personal information. This means that in practice, it does not matter whether the custodian has a relationship with the relevant data subject, or whether personal information was initially processed or obtained from the data subject by another entity. When a data custodian collects or holds information (however it comes to do so), then it will be bound by the Privacy Act. All data custodians have equal obligations under the Privacy Act.

2.1.3 The meaning of 'personal information'

The Privacy Act defines 'personal information' as:

> *'Information or an opinion about an identified individual, or an individual who is reasonably identifiable:*
>
> *a. whether the information or opinion is true or not; and*
>
> *b. whether the information or opinion is recorded in a material form or not.'*

To be personal information, information must be about a living individual (often referred to as the 'data subject' in this book).

The term 'personal information' encompasses a broad range of types of information. The definition was intended to be technologically neutral to ensure sufficient flexibility to encompass changes in information-handling practices over time. It is also comparable to the definitions used in other jurisdictions, including the definitions used in UK and EU law.

---

[5] For more information, see, for example, the eight major 'OECD Privacy Principles', which are part of the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, developed in the late 1970s and adopted in 1980 for the handling of personal data. These principles underpin many data protection laws around the world, including the Australian Privacy Act.

[6] For more information on these obligations, refer to the APP Guidelines, available on the OAIC's website at www.oaic.gov.au.

[7] Information is 'collected' for the purpose of the APPs if it is collected or obtained for inclusion in a record. This covers any instance where a data custodian obtains information, including instances where data is not directly taken from the data subjects themselves, but from a third party.

[8] An entity 'holds' information for the purposes of the Privacy Act if the entity has either 'possession' or 'control' of a record that contains the personal information.

At the conceptual level, there are two elements that information must satisfy (in no particular order) to be considered personal information. First, information must be 'about' an individual. Second, that individual must be identifiable or reasonably identifiable from the information (whether on its own, or in combination with other information that may be available). It is the second of these elements - the crucial phrase 'reasonably identifiable' – that tends to give de-identification its inherent practical complexity, as there is no exact formula for assessing when information will be reasonably identifiable. The answer to this question always depends on the context in which the information is handled.

When is information 'about' an individual?

In most cases, the question of whether information is 'about' an individual will be straightforward. This is particularly true for most unit record level data, which comprises values of attributes of a particular individual.

More broadly, for information to be 'about' an individual for the purposes of the Privacy Act, the person must be a subject matter of the information. Or put another way, the information reveals something about them. There must therefore be a connection between the individual and the information, in a way that is not too remote. Note that this will often be the case, even for information that appears to be about something else, as information can have multiple subject matters.

Data custodians should however be mindful that some data may still be 'about' individuals even where the principal subject matter appears to be about something else, and something which seems non-personal. An example of such data would be data about fires managed by a particular fire and rescue service. Despite not being directly about people, fires often happen in people's homes – and these places are in turn associated with people. So while such fire data is not primarily 'about' people, it may be about people where there is a close association between a place and a person – for example, the place is the person's address. This is because the information may reveal something about the person(s) who live there – namely that there was a fire at their place of residence.

The more difficult question usually lies with the second element of the definition of personal information - whether or not the individual is reasonably identifiable from that information. Indeed this is the principal focus of this book, which provides a method for assessing the risks of re-identification (and other disclosures) associated with particular data, and finding ways to reduce, eliminate or otherwise manage those risks so that data can be shared both legally and ethically.

The Office of the Australian Information Commissioner (OAIC)'s Guide to the meaning of personal information (OAIC 2017a) gives further examples which illustrate how the definition of personal information should be applied.

When is a data subject 'reasonably identifiable'?

The OAIC's De-identification Guidance sets out a number of factors that should be considered when answering this question (OAIC 2014c). Determining whether any data subjects are 'reasonably' identifiable in a dataset requires a contextual consideration of the particular circumstances of the case, including:

- the nature and amount of information
- who will hold and have access to the information
- the other information that is available, and
- the practicability of using that information to identify an individual.

The inclusion of the term 'reasonably' in this phrase means that where it is possible to identify an individual from available information, the next consideration is whether, objectively speaking, it is reasonable to expect that the subject of the information can or will be identified (having regard to the above factors). Even though it may be technically possible to identify an individual from information, if doing so is so impractical that there is almost no likelihood of it occurring, the information will not generally be regarded as personal information.

Therefore, an individual will be reasonably identifiable (or conversely, a de-identification process will not have been successful) where:

- it is technically possible for re-identification to occur (whether from the information itself, or in combination with other information that may be available), and
- there is a reasonable likelihood that re-identification might occur.

In many (if not all) cases where a de-identification process is undertaken, the risk of re-identification will never be totally eliminated, and re-identification will remain technically possible. This means that, in practice, whether or not de-identification has been successful will depend on whether there is a 'reasonable' likelihood of re-identification occurring.

The Privacy Act does not require de-identification to remove risk entirely, but rather requires that those sharing or disseminating data mitigate the risk of re-identification until it is very low (that is, there is no reasonable likelihood of re-identification occurring). While the risk of re-identification (identity disclosure) is the principal consideration under the Privacy Act, other disclosures, for example attribute disclosure,[9] are likely to impact on the risk of identity disclosure and so must also be considered as part of this risk assessment. Other risks, including non-legal risks, may also be posed to data subjects through the use, release or sharing of information. In addition to helping you address your legal obligations under privacy law, this book will also help data custodians to consider and take steps to minimise or eliminate these other risks.

2.1.4 Since it is not possible to completely eliminate the risk of re-identification, at what point does de-identified information become personal information?

This book acknowledges that it is not possible to draw a bright line between personal (identifiable or reasonably identifiable) and de-identified (not reasonably identifiable) information. However, the framework set out in this resource will enable organisations to assess and make decisions

[9] Attribute disclosure may not, in and of itself, constitute re-identification (and therefore, where attribute disclosure takes place, this may not necessarily mean that data was inadequately de-identified, for the purposes of the Privacy Act at least). However, where attribute disclosure might occur, this will likely have an impact on the risk of re-identification, and so attribute disclosure must be considered and protected against as part of the de-identification process. See 2.3.3 for further discussion of this.

about whether any individuals are reasonably identifiable in a dataset (and how to manage this and any other relevant risks).

'De-identified' information must carry a very low risk of re-identification having regard to all the circumstances of the particular case. There is no general rule or quantifiable threshold which applies here (although some guidance is given in Component 6 in Section 4.2).

2.1.5 De-identification is context-dependent

It is very important to remember that de-identification is not a fixed or end-state. The same information may be personal information in one situation, but de-identified information in another (see Section 3.1 and Component 1 in Section 4.1, which further outline how data situations can be dynamic, and the importance of identifying the parameters of your data situation). To use a very simple example to illustrate this point, a person's licence number (in and of itself, and in the absence of any other information) may not be personal information when in the hands of a random member of the public. However, the same number is likely to be personal information if held by an employee of the relevant motor vehicle registry, or any other person who has access to a relevant database and can look up the number and see other identifying details associated with the licence number.

The same is true for data which has undergone a de-identification process. Imagine a data custodian has undertaken a de-identification process, but they still retain a copy of the original dataset. This would enable them to re-identify the data subjects in the de-identified dataset if they wished to do so. So, the dataset may be personal information when handled by that custodian, but may be de-identified when handled by a different entity, since that is a different data situation.

Further, information may be de-identified for some parts of a data custodian's organisation, but remain personal information for others. Imagine that a data custodian decides to undertake a de-identification process on a dataset, to enable in-house research to be conducted on that data (in ways that may not otherwise be permitted if the data was subject to the Privacy Act). As for the above scenario, the custodian undertaking the de-identification will likely retain a copy of the original dataset which would enable them to re-identify the data subjects in the de-identified dataset if they wished to do so. If the custodian wants to ensure that this particular use of the dataset is de-identified and therefore outside the scope of the Privacy Act, additional controls would need to be put in place. For example, technical and/or environmental controls could be put in place to prevent those who are using the de-identified dataset for research from having access to the original dataset. In this scenario, the in-house research team may be using data that is de-identified for the purposes of the Privacy Act, while those who handle the original, identified dataset within the same organisation would still be subject to Privacy Act obligations.

2.1.6 Do privacy obligations still apply to de-identified data?

The above discussion means that in a practical sense, handling de-identified information may still carry certain 'privacy risks'. This can sometimes be difficult to conceptualise, so an example is helpful.

Imagine that a data custodian suffers a data breach, and information which was de-identified (at least within the context of that organisation) is inadvertently published on its website. If

individuals are reasonably identifiable from that published information (for example, because it could be matched with other private or public data sources), then the information would have become personal information due to the breach, and the APPs would apply in this context. This event could therefore amount to a breach of APP 11,[10] depending on the data custodian's security practices.

How should a data custodian manage such a risk, while still using de-identified data in the ways that it is legally permitted to use it? In our view, data custodians should take a risk-management approach when handling de-identified data which acknowledges that while the APPs may not apply to data that is de-identified in a specific context, the same data could become personal information in a different context. Specifically, APPs 6, 8 and 11 continue to be relevant to the handling of de-identified information, as these are the APPs which would apply if the data was to be transferred to another environment (where it could become personal information). Data custodians should therefore handle de-identified data in a way that would prevent any breaches of the APPs occurring, should such a situation arise.

For example:

- APP 6 states that personal information can be used or shared/released[11] only for the same purpose for which it was obtained, unless an exception applies.
  - APP 6 will not generally be relevant in relation to uses of de-identified information, provided that the de-identified status of the data doesn't change in the context where the use takes place. This will usually be the case where the information is confined within the context of the original data custodian's organisation. After all, the whole point of the de-identification process may have been to enable different uses of the data in-house.
  - However, APP 6 will remain relevant in relation to the sharing or release of information, if the status of the information would change to being personal information in the hands of another entity.[12]
- APP 8 prevents the sharing or release of data with an entity outside Australia unless certain steps have first been taken.
  - APP 8 will remain relevant to de-identified data, because it applies to situations where data is shared or released to an overseas entity (and the de-identified status of information may therefore change).
- APP 11 states that data custodians must take reasonable steps to protect personal information from misuse, interference, loss or 'unauthorised access, modification or disclosure' (in this context, disclosure means sharing or releasing).
  - APP 11 is probably the most important APP to keep in mind when handling de-identified data, and the data breach example given above illustrates this point. As for all valuable

[10] APP 11 imposes obligations on data custodians to take 'reasonable steps' to protect personal information from misuse, interference or loss; and unauthorised access, modification or disclosure (sharing).

[11] 'Disclosed' is the term used in APP 6. As disclosure has a specific meaning in some of the literature, we have avoided using it in this sense where possible.

[12] Data custodians will therefore needt o consider whether the information would be identifiable or reasonably identifiable in the hands of the other entity, in line with the framework outlined in this book. Otherwise, the APPs will apply to the sharing/release of that data.

> data, de-identified data should, as a matter of risk management, be protected and stored securely to prevent unauthorised access. This is particularly so in light of the obligations which will apply to data custodians from 22 February 2018 as a result of the commencement of the *Privacy Amendment (Notifiable Data Breaches) Act 2016*. [13]

To summarise, data custodians should still consider the standards that would otherwise be applied under APPs 6, 8 and 11 to data that has undergone a de-identification process. These APPs remain relevant to the handling of de-identified information - not as a matter of law (at that time), but as a matter of risk management. While this requires some extra thought and planning, it should not prevent the data custodian from using de-identified information to achieve its organisational goals.

A final point to note is that the process of de-identifying personal information can be considered a 'normal business practice' for the purposes of a data custodian's obligations under the Privacy Act, and in particular APPs 3 and 6.[14] However, a data custodian should still consider whether there may be other legal obligations, as well as potential non-legal or ethical objections to any of its proposed uses or releases of de-identified data – see Component 5 in Section 4.1 for more on this.

2.2 Why ethics is important in de-identification

It is not always immediately obvious why ethical considerations have a role to play in the process of de-identification. Most readers will understand that the processing of personal information is an ethical issue but once data is de-identified, 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 de-identification is sufficient but this might not be true from the ethical perspective. The two primary reasons why we need to consider ethics beyond the law are:

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 is an area of active debate, and this element of the framework is necessarily the most subjective. The reader may well have a different view about what is important, particularly about the 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 de-identification decision-making.

---

[13] For more information on the obligations that will apply, see the OAIC's website.

[14] Under APP 3, data custodians can only obtain personal information (whether from the data subject directly, or from any other entity) where this is reasonably necessary for, or directly related to, the entity's functions or activities (APP 3). Further, under APP 6 entities should generally only use or share personal information for the same purpose(s) that it was collected for (the primary purpose), or where an exception applies - such as where the individual would reasonably expect the APP entity to use the information for that secondary purpose. Data custodians must also describe the primary purpose of collection in their privacy policy, with a reasonable amount of specificity. In this regard, the OAIC considers that de-identification can generally be considered a normal business practice which is incidental to the primary purpose of collection. This means that as a general rule, the OAIC does not expect an entity's privacy policy to refer to de-identification as one of the primary purposes of collection of the personal information – this is implied. See (OAIC 2013).

There is growing evidence that data subjects are concerned not just about what happens with their personal information, but also about what happens with de-identified data derived from their personal information.[15] 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 de-identified – 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 (2010) 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).*

More recently, the OAIC Community Attitudes to Privacy Survey 2017 (OAIC 2017b) found that:

> *…nearly half of Australians (46%) are comfortable with government agencies using their personal details for research or policy-making purposes, four in ten are not comfortable (40%), and the balance are still unsure.*
>
> *Further, one-third (34%) of the community is comfortable with the government sharing their personal information with other government agencies. However, only one in ten (10%) is comfortable with businesses sharing their information with other organisations.*

Alternatively, an objection to data reuse might simply arise because the data subject gave their data for one purpose and you have used a de-identified version 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 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. 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:

- Are the data subjects aware that their data have been collected in the first place?
- Have the data subjects consented to the collection of their data?
- 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?
- Are the data subjects aware of the original use of their data?
- Have the data subjects consented to the original use of their data?

[15] On this point Iain Bourne from the UK's 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 a de-identified 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).

- Have the data subjects consented in general to the sharing of a de-identified version of their data?
- Are the data subjects aware of the specific organisations that you are sharing their de-identified data with?
  - Have they consented to your sharing their data with those organisations?
  - Are the data subjects aware of the particular use to which their de-identified data is being put?
  - 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 they are to take responsibility for what happens to their data and therefore the greater is your own responsibility. We shall see how this plays out in your de-identification process in Component 5 in Section 4.1.

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 filmed 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 are the implications of this discussion? 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 Terms and Conditions 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.

The key point here is that 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 crucial; it is a connecting concept which enables clearer thinking about ethics and the reuse of (de-identified) data. We will discuss 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 information 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 de-identified 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 roles and relationships between you and the proposed receiver of your de-identified 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:

- Do you (the sending organisation) have a relationship with the data subjects?
- Does the receiving organisation have a relationship with the data subjects?
- Do you and the receiving organisation work in different sectors?
- Is your organisation's area of work one where trust is operationally important (e.g. health or education)?
- 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 itself can have properties that make the data situation more or less sensitive. Three questions capture the main points here:

- Are some of the variables sensitive?[16]
- 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.
- Are the data about a sensitive topic? The topic area might be considered sensitive rather than, or as well as, the variables within the de-identified dataset because, for example, it involves particular public interest issues, or ethically challenging issues.

Again here, the more yes answers, 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 is about a vulnerable population (data sensitivity). This is shown diagrammatically in Figure 1.

[16] See the description of 'sensitive variable' in the glossary, as well as the definition of 'sensitive information', a subset of personal information, under s 6(1) of the Privacy Act.

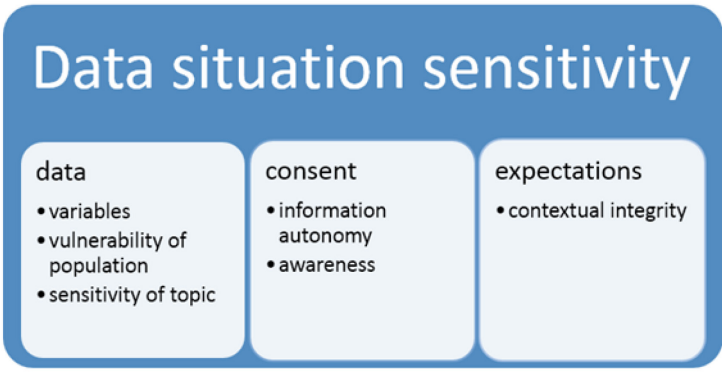


Figure 1 Diagrammatic representation of data situation sensitivity

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 loss of societal benefit 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.

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 Component 5 in Section 4.1 we will examine how it is possible to apply this in your own data situation.

3 About de-identification

In this section we first describe in a little more detail the foundations of the De-Identification Decision-Making Framework. We then give an introduction to the related Fives Safes Framework. The section closes with a brief description of the main options for making data available outside of organisational boundaries.

3.1 Functional de-identification and the data situation

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

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

We need to focus on data in its environment to address disclosure risk assessment and management because it is meaningless to attempt to assess whether data is de-identified without knowing what other information is or could be available. As we have seen, this is an important element in the application of the definition of personal information in law, and yet practitioners often attempt to judge whether data contains personal information or not using absolute criteria (the non-relative properties of the data itself). This is based in part on the misapprehension that de-identification can be absolute without causing so much damage to the data that it has no utility whatsoever and in part on being overly focused on the data itself. If de-identification is to be a useful tool for risk management, one has to specify its circumstances. Thus the only sensible response to the question 'is this personal information?' 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. **Other data**: What (other) data exist in the data environment?[17] How do they overlap with or connect to the data in question? This is what we need to know in order to identify what data (key variables) are risky, and can be used for linking one dataset with another thereby increasing the risk of disclosure.

[17] Amongst all the challenges that de-identification 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 custodians have been releasing data that has been through de-identification processes, resulting in only a small number of problems, almost all caused by inadequate de-identification 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 Component 6 in Section 4.2.

2. **Agency**: We consider agents as capable of acting on the data and in the data environment. It may seem like an obvious point but it is one worth emphasising – most data breaches involve human action or misdeed.
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 mean 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 research data centre, that is, a secure facility provided by a data custodian for researchers to access and use data onsite. 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 data centre. An environment might be defined by regulation 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 is 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 is where the concept of a data situation comes into the picture. A data situation as we have previously said can be established by mapping the flow of data from one environment to another or in more complex cases across multiple environments. What you are effectively doing is drawing a chalk line around everything in scope for consideration and which we can think about under the components of: other data, agency and structure. Regardless of the data situation type whether it be a simple share from one secure environment to another or a release of data from a secure environment to an open environment the issues for the practitioner to consider are the same (although the answers create unique data situations), they are :

- What structures inform and are likely to impact on the data situation? For example legislative and regulatory requirements, political socio-economic pressures to share data in a particular way or with particular third parties, cultures attitudes and expectations about activities, your local infrastructure and that of the data recipients.
- What other data is potentially obtainable in the environment(s) of your data situation? How easily obtainable is it? How realistic is it to be used given what you know about C.
- Who are the key agents, how might they interaction and what are the likely outcomes of those interactions.

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.[18]

In essence, the De-Identification Decision-Making Framework is a process which controls disclosure risk by considering the totality of a data situation. We give more details on this as a practical concept in Section 4. For a deeper discussion of the theory see Elliot et al (2015).

3.2 Introduction to the Five Safes

The Five Safes framework, also sometimes called the VML Security Model, was originally developed in 2003 as a way to describe the Virtual Microdata Laboratory, a 'safe centre' for confidential research data at the UK Office of National Statistics (Desai et al 2016, Ritchie 2008). The original purpose of the model was to simplify the complex discussion around data access into a set of related but independent questions, so that each topic could be dealt with succinctly and unambiguously.

The Five Safes is a framework for organising thinking about data access. The basic premise of the framework is that data access can be seen as a set of five 'risk dimensions': safe projects, safe people, safe data, safe settings, safe outputs. Each dimension provokes a question about access:

Safe projects: Is this use of the data appropriate?

Safe people: Can the researchers be trusted to use it in an appropriate manner?

Safe data: Is there a disclosure risk in the data itself?

Safe settings: Does the access facility limit unauthorised use?

Safe outputs: Are the statistical results non-disclosive?

These dimensions embody a range of values: 'safety' is a measure, not a state. For example, 'safe data' is the dimension under which the safety of the data is being assessed; it does not mean that the data is non-disclosive. Nor does it necessarily specify how the dimensions should be calibrated.

[18] Data situations can be static and dynamic, but we should also consider that data itself can be static and dynamic too. Dynamic data is 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 is moving relative to their environment. Dynamic data create a de facto dynamic data situation but so does the movement of static data.

‘Safe data’ could be classified using a statistical model of re-identification risk, or a much more subjective scale, from ‘very low’ to ‘very high’. The point is that the user has some idea of ‘more safe data’ and ‘less safe data’.

Any data access solution needs to consider all five dimensions (even if simply to note that a particular dimension is not relevant). However, each element can be reviewed independently for risk characteristics and evidence of appropriate practice.

The description of the Five Safes we present here is taken from (Ritchie and Green 2017), where it is used to break down data access mechanisms within a data access strategy into five separate components: the project purpose, the people, the settings, the level of detail in the data, and checks on any statistics produced.

3.3 The main options for data access outside organisational boundaries

The four main modes of access currently used for disseminating data for use outside of organisational boundaries are:

- Open access
- Delivered access
- On-site safe settings
- Secure virtual access

In this section we give a brief description of each of these; for more details see Appendix C.2.3.

3.3.1 Open access

Open access involves making data available with no restrictions on who can access the data, or on what they can do with it. Also, there is usually no monitoring of users or what they are doing. Often access is provided through a publicly accessible website such as data.gov.au.

Open data environments are really only appropriate to data that are either not personal in the first place or have been through an extremely robust data focussed de-identification process that ensures with a very high degree of confidence that no individual could be re-identified and no disclosure could happen under any circumstances.

3.3.2 Delivered access

Delivered access is a more restricted form of access, in which users apply to access to the data and the data is delivered to the user, most commonly through an internet portal or possibly via encrypted email. Normally the application process requires the user to specify what they are going to do with the data, and invariably the user is required to agree to specified conditions on a licence for data access.

3.3.3 On-site safe settings

The prospective data user applies for access to the data in a particular location — often in the offices of the data custodian or otherwise at a research data centre (RDC) that has been

established by the data custodian.[19] One example is the ABS DataLab. On-site safe settings are regarded as the strongest form of restricted access, usually including a high level of security infrastructure control and supervision of the user's activities. Often the user will be allowed to take away some analytical output, but usually only after it has been checked by output checkers for disclosure risk.

3.3.4 Secure 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 access to a copy of the data from one's desktop.[20] There are two variants on the virtual access theme: direct virtual 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. Often, output is checked in the same manner as in an on-site safe setting.

Analysis servers go one step further than direct virtual access 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 disclosure risk.

[19] Examples are the ABS DataLab (http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL)), Statistics New Zealand Data Lab (http://www.stats.govt.nz/tools_and_services/microdata-access/data-lab.aspx), the Administrative Data Research Centres in the UK (www.adrn.ac.uk) and the US Federal Statistical Research Data Centres (http://www.census.gov/about/adrm/fsrdc/locations.html).

[20] An intermediate hybrid approach is where safe rooms 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 by the ABS as a method for allowing academic researchers to have access to administrative data within Australia.

4 The De-Identification Decision-Making Framework

The DDF is made up of ten components, and we will describe each of these components in detail in this section.

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

Presentation of the DDF components as a numbered list is not intended to suggest that this is the right order to do them, or that you can run through them in a linear fashion and tick off as you go down the page. For example, stakeholder engagement plays an important role in shaping purpose of the data access arrangements, costs, breach policies, outputs and public messaging.

The ten components of the DDF are grouped into three core de-identification 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 it 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 use 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.

Figure 2 gives a diagrammatic representation of the De-Identification Decision-Making Framework.

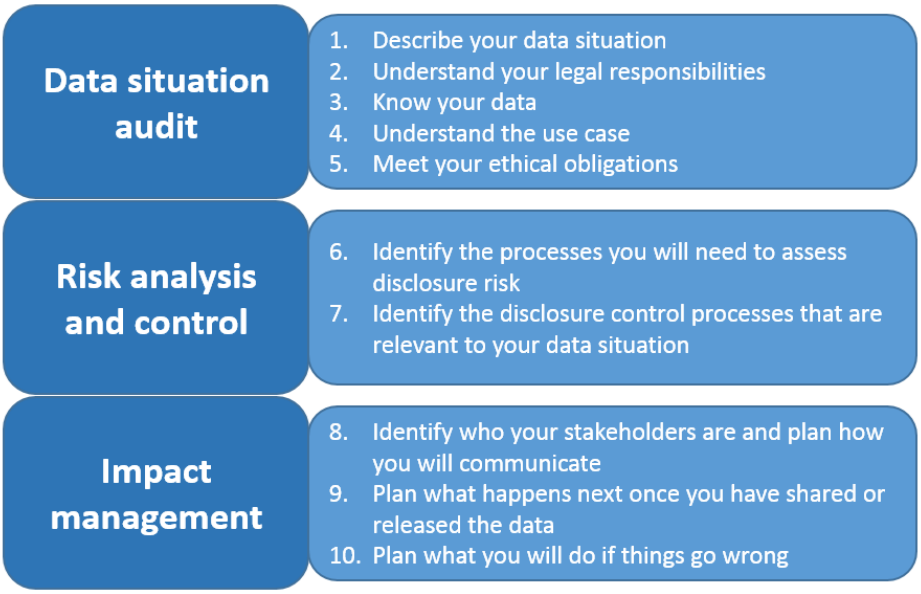


Figure 2 Diagrammatic representation of the De-Identification Decision-Making Framework

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-building purposes, to understand how a privacy breach might occur and explore its possible consequences, or to develop a sound understanding of the important issues in the de-identification process. Others might use it directly to support their management of the risk of a privacy breach, to reduce it to a very low level.

De-identification is not an exact science and, even using the DDF at this level, you will not be able to avoid the need for complex judgement calls about when data is sufficiently de-identified given your data situation. The DDF will help you in making sound decisions based on best practice, but it is not a step-by-step 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 de-identification process, particularly with the more technical risk analysis and control activities. However, even when working with an expert, the DDF 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 DDF 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 DDF to expand your knowledge or to support decision-making, it is important to recognise that it is not a simple checklist 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.

4.1 The data situation audit

The data situation audit is essentially a framing tool for understanding the relationship between your data and its environment, which should help you to scope the de-identification process appropriately for you to share or release your data safely. It will help you to clarify the goals of the process and will enable the more technical aspects of the de-identification process (Components 6 and 7 of the DDF) to be planned and conducted more rigorously.

Component 1: Describe your data situation

In Section 1 we introduced the term data situation to refer to the relationship between data and its environment. So for example, your organisation itself will constitute an environment, while any proposed share or release would involve another environment. These environments will have different configurations of the same core features: people, other data, infrastructure, and governance structures. Figure 3 gives a diagrammatic representation of a data situation.

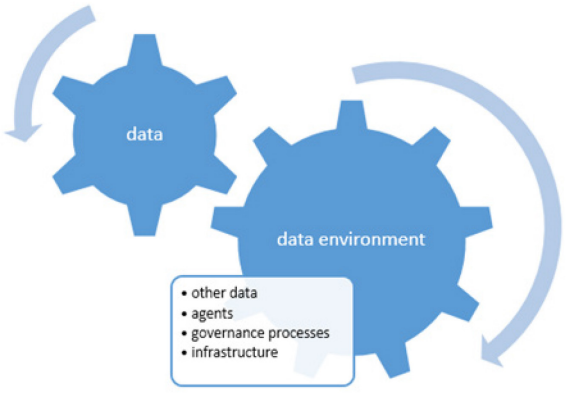


Figure 3 A data situation involves the relationship between data and its environment, where the environment comprises people, other data, infrastructure, and governance structures

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 sharing or releasing processes take place within dynamic data situations in which data is intentionally moved from one environment to another. A dynamic data situation might be relatively straightforward, simply involving the movement of data from just one environment to another environment. Often though, it may involve multiple environments.[21]

At this stage we want to familiarise you with the concept of data moving between environments. While 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

[21] 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.

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.[22]

We 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 has been subject to a de-identification process.

Imagine that PubT (a franchised public transport provider) collects data containing personal information from its customers relating to public transport usage. PubT plans to share a de-identified version of the data with the Local Council Area of Anycouncil, which wants to use it to support (better) provision of public transport. PubT de-identifies the data by removing the direct identifiers, i.e. the customers' names and addresses, and unique reference numbers like tax file numbers, and by aggregating the detail on several key variables. However, it leaves some key variables – which are of particular interest to Anycouncil - unchanged. Call PubT Environment 1.

Anycouncil signs a contract with PubT which

- enables Anycouncil to analyse the data (for a specified purpose other than that for which it was originally collected),
- prohibits Anycouncil 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
- requires Anycouncil to keep the data securely, and to safely destroy it once it has finished using it. After this contract is signed, the de-identified dataset is passed to Anycouncil, so call Anycouncil's arrangements Environment 2.

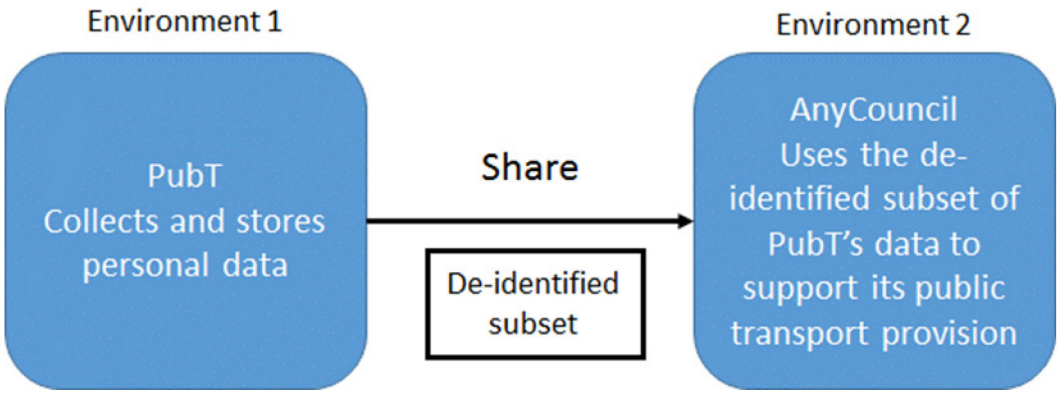


Figure 4 Data flow between two environments

Figure 4 illustrates the intentional movement of data from Environment 1 to Environment 2. The data flow between PubT and Anycouncil 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 transferred data within Anycouncil's environment is considered low risk even though it contains some detailed key variables. This is because the

[22] Conducting a Privacy Impact Assessment (PIA) may assist with mapping the relevant information flows. See (OAIC 2014b).

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 inappropriately de-identified datasets such as the Netflix, AOL, and the New York taxi driver open datasets.

Data situation: simple share with secondary open release

Consider this example further and imagine that Anycouncil 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 third environment in the chain is the open environment. The new picture of the data flow is shown in Figure 5.

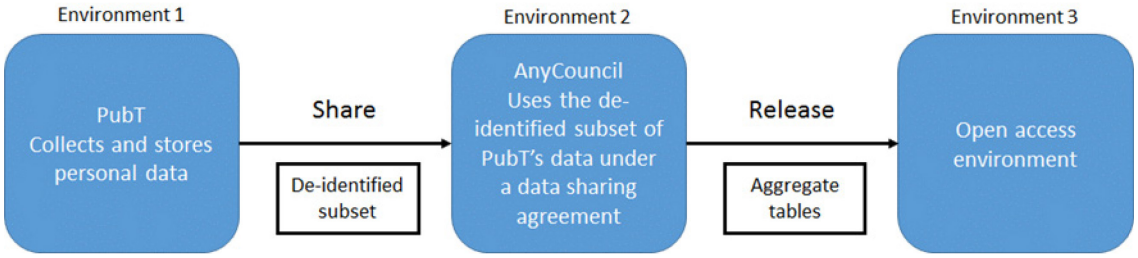


Figure 5 Data flow between multiple environments

The data flow between PubT, Anycouncil and the open access environment defines the parameters of Anycouncil's data situation for the de-identified public transport data.

Anycouncil, as stipulated in its contract with PubT, cannot release the de-identified data given to it in its original form without permission from PubT, due to contractual arrangements. Prior to releasing any data, Anycouncil should carry out a disclosure risk audit of the open access environment[23] and then further de-identify the intended disseminated data product as necessary given the likely use case(s) (this is covered in Component 4 of the DDF). Failure to adequately de-identify the output data could result in a breach of APP 6 when the data is shared or published.

Data situation: simple share with secondary controlled release

In this example we look at the data flow across environments involving personal data and data with direct identifiers replaced with random codes. The situation is represented in Figure 6. Imagine that Anycouncil collects public health data for its area. It has statutory powers to share (some of) the public health data with a government department (GD) to support its work on health promotion and disease prevention. Both Anycouncil and GD are data custodians. We call Anycouncil Environment A.

[23] In Component 6 of the DDF 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.

The GD as part of its role in health promotion (and in accordance with its agreement with Anycouncil) creates a subset of the data, replaces the direct identifiers with random codes, then makes it available within a secure setting for reuse by approved accredited researchers. The GD is Environment B.

The secure setting is designed to ensure that the data, when in that environment, is functionally de-identified. It places restrictions on who can access the data, on how it can be accessed, and on what auxiliary information can be brought in and out of the secure laboratory environment. The secure laboratory is Environment C.

An approved accredited researcher carries out an analysis in the secure laboratory producing statistical output, such as regression model output, that will be included in any publication arising from the research. This output is first checked by secure laboratory staff to ensure that it is not disclosive, so it is passed as 'safe'. The researcher duly writes up and openly publishes the results, which contain some of the statistical output. The publication of the research is a fourth environment, which we call Environment D.

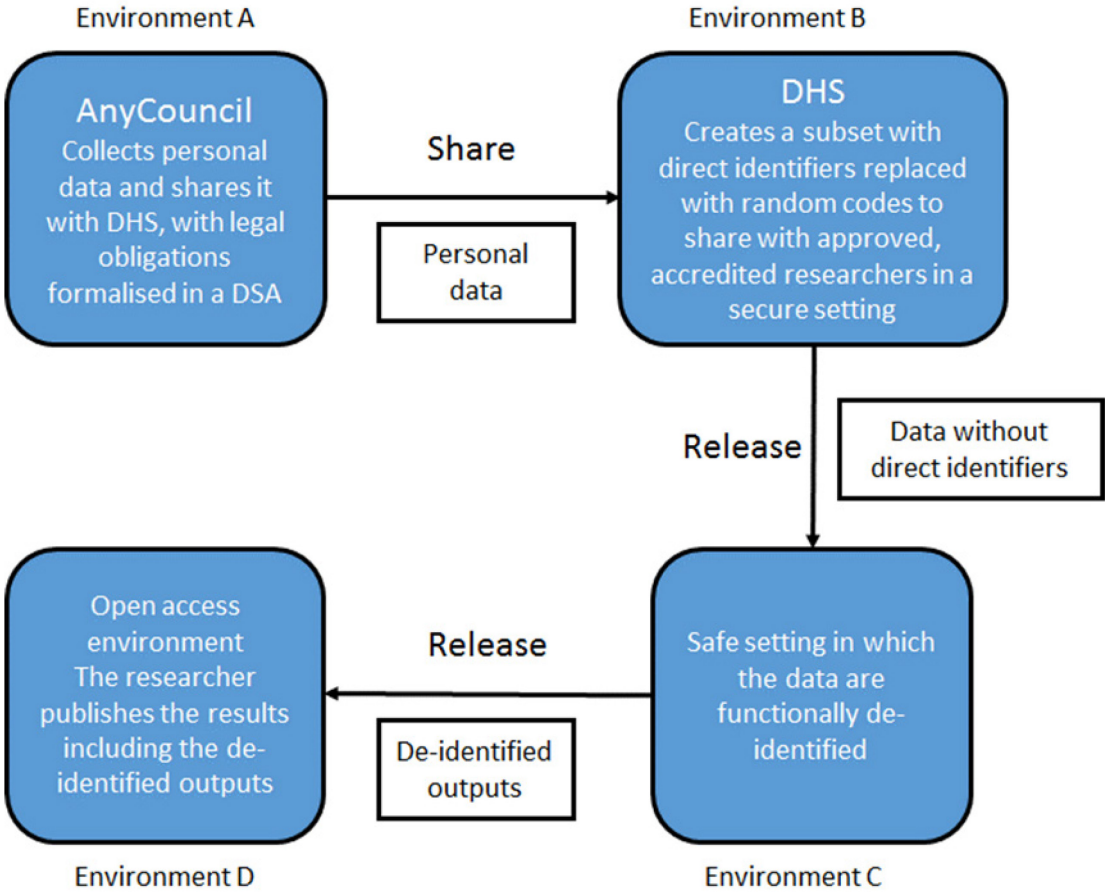


Figure 6 Movement of data across data environments

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 de-identified (for example the data with direct identifiers replaced by random codes 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 statistical output has to be checked and verified as 'safe' by the secure laboratory before the outputs can be removed from the secure lab.

It is worth stressing that while a data situation might be complex, it should not make you feel that it is 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 DDF, but it need not be the starting position. You should not lose sight of the enormous range of benefits that can and do come from sharing and releasing data.

Component 2: Understand your legal responsibilities

Now that we have explored the concept of a data situation, we can move on to review the legal obligations which will apply as data moves between different environments. In this regard, you should consider your own applicable legislation, as well as the matters discussed earlier in Section 2.1, and most importantly the status of the data in each particular data environment. The key questions are:

1. is the data personal information or de-identified data, and

2. if it is de-identified data, what controls need to be in place to maintain this status?

Where data is personal information

If your data contains personal information, you should establish how you will comply with your obligations under the APPs when handling that data; for example, whether you have authority to obtain the information in the first place (under APP 3), and whether your intended uses of the data are permitted under APP 6. Data custodians should be aware that the APPs will apply to entities that 'hold' or 'collect' personal information for the purposes of the Privacy Act. All data custodians have the same obligations under the APPs, whether or not they are handling data on behalf of someone else, or are only handling data very briefly (for example, where they are transmitting and/or temporarily storing data).

The OAIC has produced significant guidance which can help you to understand and comply with your general APP obligations (OAIC 2013). All data custodians should be aware that breaches of the Privacy Act or obligations under the APPs may be subject to regulatory action by the Information Commissioner (OAIC 2015).

Where data is de-identified

If the data is de-identified, you will need to consider how to ensure that it remains de-identified when in your possession. In particular, where environment-based controls have been used, data custodians will need to ensure that they maintain the data's de-identified status by complying with any of the requirements which have been imposed by the 'original' data custodian (as a condition of them being able to receive and use the data).

As discussed in 2.1, while the APPs do not apply to de-identified data, a change in the data environment may result in such data becoming personal information once again. This means that data custodians should handle de-identified data in a way that would prevent any breaches of the APPs occurring, should such a situation arise.

All data custodians should also be mindful that where de-identified information becomes personal information once again (as a result of a change of circumstances, a data breach, or any other reason), the APPs will apply to that information, and the data custodian could be subject to

regulatory action by the Information Commissioner in the case of any breaches of the Privacy Act (OAIC 2015).

Checklist for understanding of legal responsibilities

These issues have already been explored in 2.1, however they are summarised here briefly by way of a checklist:

1. Is my data personal information? The answer will be 'yes' if:
   - the data is about individuals, and
   - those individuals are identifiable or reasonably identifiable.
2. If my data is de-identified, can I forget about my privacy obligations?
   - De-identified data is not subject to the APPs.
   - However, data custodians should ensure that de-identified information stays de-identified when in their possession and the risk of re-identification remains acceptably 'low' in the particular context.
   - Further, you must still consider, as a matter of risk management, which APP obligations should still be adhered to (in particular, APPs 6, 8, and 11) in case your de-identified data becomes personal information in another context. This is particularly relevant where environment-based controls have been used.

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 itself. 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 Components 6 and 7 of the DDF. 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 de-identification, whilst other points are useful indicators as to the general level of risk[24] 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.?

[24] 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 de-identification.

Data subjects

In most cases, who the data is about is a straightforward question. However, as we demonstrated in Section 2.1.3, data can be indirectly about people when they are directly about something else (in the example we used data about 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. because it contains their professional opinion, and also reveals other information about them).

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 is commonly stored in digital datasets 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 a de-identified version of the microdata, but instead to aggregate your data and make it available as a de-identified 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 small census output areas (for example ABS Statistical Area Level 1 (SA1) areas) or postcodes, attribute disclosure 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, identity disclosure is a particularly challenging problem. This is because of the difficulty of determining which variables or combination of 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 section).

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.

At this stage you are not attempting to form fully specified scenarios but simply to 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 de-identification problem.

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 tax file numbers and Individual Healthcare Identifiers. In most dynamic data situations you will suppress (or possibly replace with

random codes or pseudonyms) the direct identifiers as a first step. Therefore, being clear about which variables are direct identifiers is important.

An indirect identifier is any variable that can be used to identify an individual with a high probability, either alone or together with other indirect identifiers, and in combination with auxiliary information. Almost any variable can be an indirect identifier, depending on the auxiliary information available to the intruder. 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 Tasmania 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 re-identifying them.

A more common scenario involving indirect identifiers is that the intruder has access to auxiliary information in the form of a dataset that contains the indirect identifier(s) together with some direct identifier(s) such as name and address, thus the indirect identifier(s) plus the auxiliary information re-identifies the individual. For example, even a number such as body mass index (BMI) can be an indirect identifier if the intruder has a register of individuals and their BMI (such as for a weight loss program), and there are individuals with unique values of BMI. Examples of such indirect identifiers include: sex, date of birth or age, location (such as post code, census geography, proximity to known or unique landmarks), language spoken at home, ethnic origin, Aboriginal and Torres Strait Islander identity, total years of schooling, marital status, criminal history, total income, visible minority status, disability, profession, event date (such as admission, discharge, procedure, death, specimen collection, visit/encounter), code (such as diagnosis code, procedure code, adverse event code), country of birth, birth weight, and birth plurality. When in doubt, it is safest to assume a given variable is an indirect identifier.

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).

Sensitive variables: Sensitive variables are thought to increase re-identification risk because

1. they are more likely to be targeted because they are interesting, and

2. the impact (and potential harm) of a disclosure may be greater.

'Sensitive information' encompasses a wide range of information and holds a special position in the Privacy Act. Within the Privacy Act, the definition of 'sensitive' is based on a list of categories which may be regarded as an incomplete list, if one considers what might intuitively be covered (financial and credit data, for example, is notably absent). So you may consider that there are other categories of data which are sensitive and identifying, and take that into consideration when making decisions about the risks associated with handling such data in a particular data environment. For example, 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 variables 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 information is also likely to be more damaging to the data custodian than disclosure of non-sensitive variables, 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 a de-identified 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.

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 depend on the type of the data (such as survey, transactional, or administrative data) as well as other factors. Relevant key data properties can 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 custodian 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, the harder it is to identify people correctly from it.[25] 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 Appendix C.2.2.

Hierarchical data – data about groups as well as individuals: This is data that contains information for members of a group (such as members of a household, or data linked to locations) who are linked with one another. The data is 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).

[25] 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.

Time-stamped (or longitudinal) data: This is data about a defined population which is collected over time. A common example is transactional data such as supermarket purchases over time linked to a loyalty card. 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 therefore identifiable or reasonably identifiable.[26]

Population or sample data: Population data is data that represents 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 birth. 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 release and what is the most appropriate means by which to do this.

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

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

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

There are numerous reasons for sharing or releasing 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 a Freedom of Information (FOI) request. Thinking through why you are disseminating your data automatically brings in the other two questions, the 'who' and the 'how' of access.

[26] Analytically, longitudinal data could be treated as a slow form of dynamic data (data that updates over time). In practice however it is analysed as a static dataset 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.

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 channels.[27]

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 scope of this section 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.

That there will be some benefit to the reuse of data is widely accepted 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 itself. 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

1. makes clear who is responsible for what and

2. ensures fair processing, usage and retention of the data.

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

By using a contract or 'Data Sharing Agreement' (DSA) you can manage (some of) the disclosure risk associated with the data share.

Data release options: See Section 3.3 for a summary of this topic, and Appendix C.2.3 for more details. 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, its sensitivity, and the proposed usage. 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 itself. Allowing greater access to your data does not automatically produce high risk, as long as your data is sufficiently de-identified given the release environment.

It is worth noting that in applying the risk-utility concept (Appendix C.1) you will need to think beyond the impact of de-identification techniques on your data. Think also about what 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 the level of data modification required to achieve de-identification, always ensuring legal compliance including considering the intrusiveness and cost of any approaches used. This balance is a well-recognised trade-off and underpins the challenging task of producing safe, useful data.

Component 5: Meet your ethical obligations

As discussed in Section 2.2, 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 de-identified data.

Consent and other means

Where possible, seek consent from data subjects for what you intend to do with their data once de-identified. As we stressed in Section 2.2, consent is not a panacea. However, 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 assurances you give.

When consent has not been sought, you should aim to be as transparent as possible, and engage with stakeholders where practical.

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, in that it is 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. One place to start would be the OAIC's relevant guidance (OAIC 2017b).

The importance of good governance

Key to ensuring you meet your ethical responsibilities as a data custodian is good governance. On a broad level governance is about the organisation of your data processing activities as 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 your processing responsibilities.

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 de-identification process and ensure that they have the necessary skills and knowledge.
- 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 custodian.
- 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 privacy 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 (see OAIC 2014b).
- Establish principles, policies and procedures for identifying and dealing with cases where de-identification may be problematic to achieve. You should also consider at what point in the process (in dealing with a difficult case) you should seek guidance and advice published by bodies such as the ABS, National Statistical Service (NSS)[28] and OAIC.

[28] http://www.nss.gov.au

- 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.

Data breaches

1. Familiarise yourself with the requirements of the Notifiable Data Breaches scheme, and in particular the definition of eligible data breach, which will apply from 22 February 2018.[29]
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. See (OAIC 2015) for useful advice in relation to these matters.

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 de-identification. For example you could regularly view the OAIC website for information on de-identification and data protection for personal data, and the ABS and NSS websites for information on de-identification and anonymisation more generally.
- Talking to other organisations in your sector to share best practice. You might want to consider going to events such as the OAIC's annual Privacy Awareness Week events to keep up to date with current issues and networking with other people working in the privacy field.

[29] A data breach can generally be defined as 'unauthorised access to, unauthorised disclosure of, or loss of, personal information held by an entity': see s 26WAE(2) of the *Privacy Amendment (Notifiable Data Breaches) Act 2017* (which becomes part of the Privacy Act on 22 February 2018). While data breach is usually used as a general term, 'eligible' data breaches for the purposes of the Privacy Act occur only 'where the access, disclosure or loss is likely to result in serious harm to any of the individuals to whom the information relates'.

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 Privacy Act to undertake a Privacy Impact Assessment (PIA),[30] there are many good reasons for having one when you process data. It will

1. help you be aware of and address any particular privacy issues

2. ensure the transparency of your activities,

3. promote trust in what you do, and

4. help you to comply with the Privacy Act and any other relevant legislation.

For further information see the OAIC's Guide to undertaking privacy impact assessments (OAIC 2014b).

Figure 7 shows a diagrammatic representation of the data situation audit, the first part of the De-Identification Decision-Making Framework.

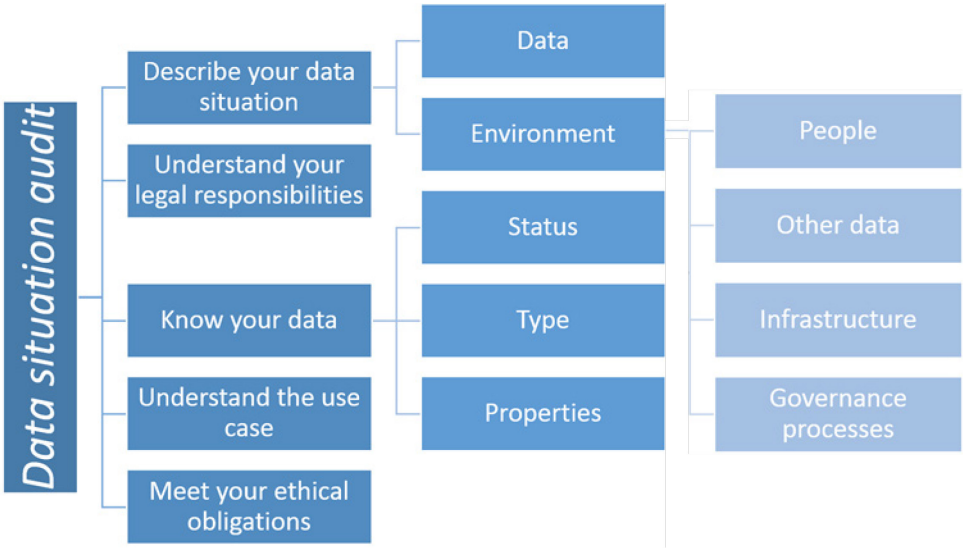


Figure 7 Diagrammatic representation of the data situation audit, the first part of the De-Identification Decision-Making Framework.

4.2 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 time and budget are limited).

[30] However, note that the *APS Privacy Governance Code*, slated for introduction in July 2018, may require Australian Government agencies to conduct PIAs for particular projects. See the OAIC's website for further updates on the introduction of this Code.

Risk assessment and control 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, while complying with legal and ethical obligations. The use of resources to address potential problems should be proportionate to the risks presented by the data - principally, the risk of re-identification or other disclosure (but also other risks that could arise from a breach or a misuse of the data).

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:

1. determine whether your data should be shared or released at all;
2. determine how much disclosure control should be applied; and
3. 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 de-identification 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 need to 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 not possible to predict what information might become publicly or privately available in the future, from sources other than yourself, which could be linked with your data, resulting in re-identification. 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. We recommend 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 3 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	Lower data quality reduces the risk
Age of data	Older data is 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	Key variables form the basis of any attack
Microdata	Re-identification disclosure is a particular problem
Aggregate data	Attribute disclosure and disclosure by differencing are particular problems

Table 1 Risk-relevant features

This top level analysis enables you to identify where you need to 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 is 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 modifications of 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.[31] Initially, removing low-utility/high-risk features will not impact on the overall utility. However, eventually 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

The purpose of scenario analysis is to ground your assessment of risk in a framework of plausible events. If you use the framework outlined in Appendix B.2 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

[31] One factor to bear in mind here is the nature of sharing/release 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.

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 title registers, real estate agents' lists, newspaper reports, archived reports and announcements, local government records and vital statistics such as birth, death and marriage registry 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 and professional networking sites). Needless to say this is a growing source of publicly available information and Elliot et al (2016b) 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.[32]
- **Official data releases**: including data releases from the Australian Bureau of Statistics, 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 is 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 is 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 is not safe relative to these standards, then in itself that indicates that you

[32] Ideally if multiple organisations were releasing open data on the same population then they would co-ordinate their de-identification 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 circumstances will usually only arise when the organisations are closely related. This, in principle, allows key extension; as both datasets are de-identified we are not here talking about direct re- identification, but the fusion of two de-identified datasets could make both more vulnerable.

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.

For example, imagine that you are a government agency wanting to release a dataset of human services care users as open data. Suppose the dataset contains 7 variables: age (banded), sex, cultural and ethnic group, postcode, 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 (1999) 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 linking.
- **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**:
  - Postcode
  - 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 2016b 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 interpreting their findings and recommendations. In this section, we will set out a process that could be performed in-house, without (i.e. before) consulting de-identification 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 puzzle in itself. There are three key questions whose answers will guide you:

1. Is your data a sample or a population?
2. If it is a sample then is it (approximately) a random sample?
3. Does your scenario involve identifying a given individual in the data, or attempting to re-identify anyone at all in the data?
4. Does your scenario assume response knowledge and if so at what level?

By 'population' here we do not just mean the Australian 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 is about all the members of

the Anytown 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 Anytown Cycle Club members all have to live in Anytown 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:

1. Give up now and do not release/share the data
2. Proceed to apply disclosure control (see Component 7 below), or
3. If you still want to persist with your proposed release/share then you will need to carry out penetration testing (which we discuss shortly).

If you go for option 2 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)[33] 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.[34] 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 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 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 is non-sensitive, and coupled with that an indication of the type of environmental solution that might be suitable. If the data is 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 is 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 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.

[33] 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.

[34] 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.

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 an 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 its 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:

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 increasing the likelihood of both false positive and false negative matches.[35]

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 is 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 eight of them.[36] 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

[35] 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).

[36] 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.

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 principle 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[37] 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 the underlying records are likely to be interesting and unusual (though probably not as unusual as the ones you identify in stage 1).
3. Repeat step 2 with the 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. At this point it may be useful to ask your colleagues to check and provide comments on your list. 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.

---

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

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 modifying 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 of the original UK version of this book 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 useful 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 test:

1. data gathering

2. data preparation and harmonisation

3. the attack itself, and

4. 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 penetration test reported in Elliot et al (2016b) gathered information on 100 individuals, and took about three person-months of effort. That test also included a second augmented attack using data purchased from a commercial data broker.

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 custodian will provide the linker with a small sample of random direct identifiers (usually name and residential address), drawn from the dataset. If not then the simulated linker 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 has 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 a dataset with identifiers 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 will still be 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 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 matches between your auxiliary 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 shows what an output from this process might look like.

Name	Address	Record No	Confidence	Effective Confidence
Johnny Blue	10 Callistemon Gardens….	10985	95%	95%
Jamie Green	68 Zieria Walk….	45678	95%	95%
William Pink	53 Verticordia Lane….	42356	90%	60%
Fred Purple	39 Verticordia Street….			30%
Archibald Black	68 Callistemon Walk….	671	85%	85%
Jane Indigo	23 Rhododendron Gardens….	37	80%	40%
		9985		40%
Patricia Vermilion	20 Verticordia Drive….	70637	60%	60%

Name	Address	Record No	Confidence	Effective Confidence
Wilma White	53 Lambertia Drive….	68920	50%	50%
Gertrude Gold	57 Pittosporum Street….	35549	40%	40%
Brittany Magnolia	12 Acacia Walk….	22008	30%	30%
Petra Puce	75 Callistemon Street….	68680	30%	30%
Stephanie Red	11 Pittosporum Drive….	81994	30%	30%
Simon Violet	136 Acacia Street….	91293	20%	20%
Estimated number of correct matches			7.05	

Table 3 An example output from a penetration test

We see from Table 3 that there are two individuals matched against record 42356 and that the individual 'Jane Indigo' is matched against two records. Here the linker 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, they 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).

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

- the dataset contains data for a person who is highly similar to our target – their statistical twin – or
- 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.[38]

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 linking. If the linker 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 (2016a) list the following:

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

[38] While incorrect information or opinion about an identifiable individual is still personal information for the purposes of the Privacy Act, (correct or incorrect) information that has been incorrectly linked to an identifiable individual is not. This is because both the 'about' and identifiable elements of the personal information definition need to relate to the same person.

2. **Expertise variance**. Typically the linker will be an expert or at least skilled and knowledgeable about data. Even if they simplify their linkage 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, penetration 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. **Penetration 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 penetration 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 is 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 broad level of successful matching poses an unacceptable risk? It is difficult to give a general answer to this question. But if you have produced a table like Table 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. As discussed above, this is in part because the possibility of causing serious harm can increase the likelihood that an intruder would be motivated to carry out an attack.

Related to this is the importance of cross-checking the correct match rate achieved against the rate estimated by the linker. To derive the former, simply sum up the confidences (converted to proportions). As you can see in Table 3, the estimated number of correct matches is 7.05. If your number of correct matches varies significantly from this estimated figure then the linker 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 used? The methods plus the parameters used? This will partly depend on the moment in the de-identification 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 modification 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 its environment. If your risk analysis in Component 6 suggests that you need stronger controls then you have two (non- exclusive) choices:

1. Reconfigure the data environment

2. Modify the data, including possibly reducing the amount of data under consideration

In Appendix C we describe a range of disclosure control techniques, and their pros and cons. You can find further information on controls for both tabular data and microdata in the ABS Confidentiality Series on the NSS website.

Figure 8 shows a diagrammatic representation of disclosure risk assessment and control, the second part of the De-Identification Decision-Making Framework.

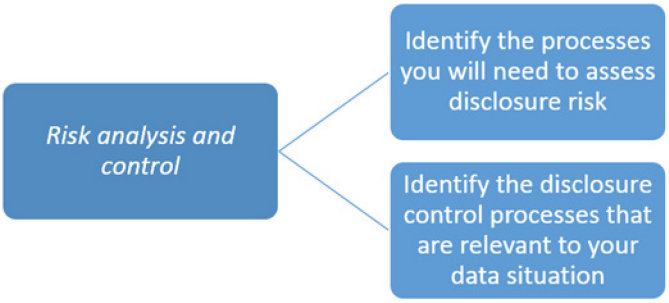


Figure 8 A diagrammatic representation of disclosure risk assessment and control, the second part of the De-Identification Decision-Making Framework

Reconfiguring the environment

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 de-identification 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 de-identification, at least in part, through reconfiguring the environment. Some of the options to consider are:

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

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. Such controls can be especially useful if the data is sensitive.

Modifying 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.

- Keeping the use case in mind, can you lose detail on your key variables to reduce the measurable risk?
- 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 loses most of its 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 sampling 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 use cases involving release, you should give some serious consideration to this possibility.

One overarching advantage of data reduction methods, 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 modification 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 modification 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 modification.

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

Now you have to decide whether the modification should be random or targeted. Random modification in fact has relatively low impact on the risk –you will have to do quite a lot of modification before you get a significant impact. Random modification works by reducing the baseline confidence in any match. Targeted modification 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 reduce variability and introduce bias. So our guidance is to do this only very sparingly.

The second issue is that once you have modified the data then the standard risk metrics will no longer work. There are techniques for measuring post-modification risk, but these are experimental and complicated to implement. So there are two options:

1. you add in modification to pick up a small amount of residual targeted risk when you are quite close to your acceptable level anyway or
2. you carry out a penetration test.

In general, the application of data modification tools can be tricky, so it is worthwhile to first consider how far you can go with just environment-based controls – without, of course, introducing an overly restrictive setting. Then data modification tools can be used for the balance of the protection – with the advice and/or guidance of an expert.

4.3 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. The OAIC's Data breach notification — A guide to handling personal information security breaches is also a useful resource in this regard (OAIC 2014a).

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, its use and reuse, whether privacy is a high priority in your organisation, and whether assurances of protection 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, special interest groups and the media may also want to know about the **how** of your processing activities, such as how are they de-identified 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 feedback), 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:

- **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.
- **A web or mail survey:** You could develop a short 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.

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 privacy 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 Australia will include groups such as the Department of Health, health-related peak bodies, GPs, allied health professionals and other clinicians, patient and health consumer organisations, community and other support organisations, 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:

1. to promote trust in your organisation's handling of data
2. to build relations with relevant specialist groups, and
3. 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 the protection of privacy and robust de-identification practice.
  - Objective 2: Produce a report on your data sharing 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.[39]
- **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. Further, 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, regardless of your own actions, will also be scrutinised by your stakeholders and others very closely for a period.

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

Having shared or released a dataset that has been de-identified for a given data situation, 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 into which you have shared or released your data to change. So whilst your data may be considered to be de-identified at the time of its release, this may not remain the case in the future. Of course any risk assessment needs to at least consider what might happen to data and its environment in the future – however changes may occur that are not foreseeable. Further, your obligations under the Privacy Act continue to apply into the future, even if no further action is taken in relation to that data from the data custodian's point of view.

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

[39] 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/

1. Keeping a register of all the data you have shared or released, including a description of the associated data environment(s).
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.
3. Be aware of changes in the data environment and how these may impact on your data. This means
   - 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;
   - monitoring changes in the law or guidance on data sharing and dissemination by engaging with relevant organisations such as the OAIC, and
   - 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 best practice, things can go wrong. As identified in Component 2, it is important that you have effective governance policies and procedures in place which 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. Again, the OAIC's Data breach notification — Guide to handling personal information security breaches can help with this (OAIC 2014a). The OAIC will also be providing updated guidance in due course on the data breach notification scheme, which will enter into force on 22 February 2018.

Ensure you have a robust audit trail

Being able to provide a clear audit trail which takes into account all relevant de-identification activities and processes will be crucial for the purposes of:

1. demonstrating that you have followed all correct procedures, and
2. 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 de-identification process may, however, impact on utility and for this reason you may wish to provide a top level public narrative about your de-identification 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 de-identification 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 de-identification issues and techniques, data security, privacy law etc.

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

- Organising regular team meeting/briefings to look at de-identification issues such as ‘what are my responsibilities under a Data Sharing Agreement when processing data from another source?’
- Implementing a staff non-disclosure agreement to provide clear guidance to staff about their privacy and non-disclosure obligations 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, this 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 OAIC, 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:[40]

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.

If there are reasonable grounds to suspect that there may have been an eligible data breach, from 22 February 2018, you must carry out an assessment of whether an eligible data breach has occurred.

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 estimate 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?
- Is the data sensitive?
- Is anyone, and if so how many, likely to be affected by the breach?
- What is the nature of the harm likely to be experienced?
- Finally, is the breach an 'eligible data breach' under the Privacy Act? That is, is the breach likely to result in serious harm to any of the individuals to whom the information relates? If so, then specific procedures may need to be followed (see the OAIC's Guidance, OAIC 2014a).

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.

---

[40] For more information, see (OAIC 2014a).

Step 4: Notify the appropriate people.

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

If the breach is an 'eligible data breach' under the Privacy Act then you must notify affected the OAIC and affected individuals (as soon as practicable). You should also consider whether to voluntarily notify affected individuals and the OAIC, even if the data breach isn't an eligible data breach.

Step 5: Penalties

Include in the plan details about any penalties associated with behaviours of staff or other users indirectly or directly leading to a breach. Make sure identified penalties are fair, consistent and enforceable. It may also be worth considering penalties that may be imposed by the OAIC as the regulator.

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.

In addition to the general advice given above, data custodians should keep abreast of any new guidance to be provided by the OAIC, following the introduction of the notifiable data breach scheme.

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.

Figure 9 shows a diagrammatic representation of impact management, the third part of the De-Identification Decision-Making Framework.

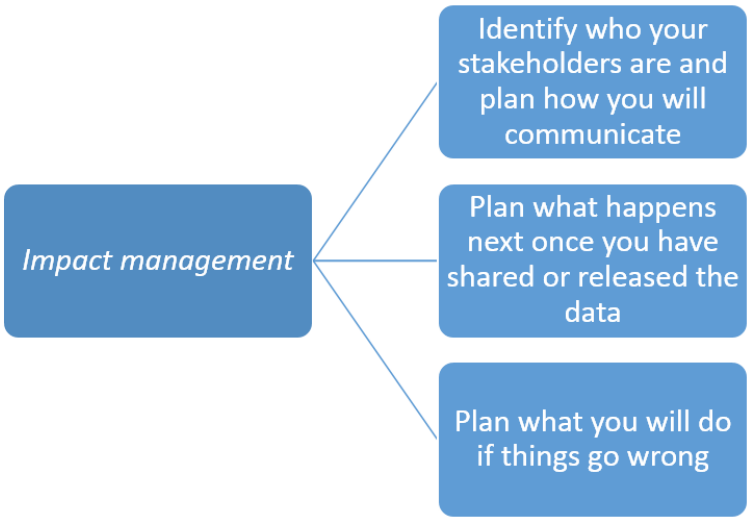


Figure 9 Diagrammatic representation of impact management, the third part of the De-Identification Decision-Making Framework.

4.4 Closing remarks

In this section, we have described the De-Identification 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 it does provide a structure for your decision-making, which can reduce the complexity of the process of de-identifying your dataset before you release or share it. Each of the components in the framework will require you to think and explore, and to incorporate the outcomes into your de-identification decision-making. With adequate time and resourcing, we hope that the framework can assist in transforming data management and sharing from a daunting and confusing process, into one which allows you to practically optimise the utility of your data.

4.4.1 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 in the Springer Lecture Notes in Computer Science series every two years since 2004.

For treatments of the end to end de-identification problem that focus particularly on health data we would recommend the reader looks at the work of Khaled El Emam 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 de-identification 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 de-identify their own data and possibly serves as a warning of what is likely to happen if data custodians do not improve their privacy practices (Brunton and Nissenbaum 2015).

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 important issue of whether consent can ever truly be informed.

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 this topic. 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, and this body of literature raises a number of very interesting questions for further reflection.

4.4.2 Next steps for the framework

We regard this framework as a changing and adapting open document. The data environment is constantly changing and new forms of data are appearing all the time. Therefore, we intend to review the framework and revise it on a regular basis.

Abbreviations

ABS	Australian Bureau of Statistics
APP	Australian Privacy Principle
DDF	De-Identification Decision-Making Framework
EU	European Union
ICO	UK Information Commissioner's Office
NSS	National Statistical Service
OAIC	Office of the Australian Information Commissioner
ONS	UK Office of National Statistics
PHRN	Population Health Research Network
SDC	Statistical Disclosure Control
UK	United Kingdom
UKAN	UK Anonymisation Network

Glossary

Attribute disclosure (Attribution): This is the process of associating a particular piece of data with a particular population unit (person, household, business or other entity). In essence, it means that something new is learned about that population unit. Attribute disclosure often follows re-identification, however it can also occur without re-identification.

Auxiliary information: Information, usually in the form of a dataset, that is available to the intruder and is not contained within the target dataset.

Data custodian: An entity which handles data. For the purposes of this document, we assume that data custodians are handling data that contains (or is derived from) personal information.

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 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 linkage: A process that compares records from one or more datasets with the objective of identifying pairs of records that correspond to the same population unit. Such pairs of records are said to be `matched'. This is also called statistical linkage, data linkage, or record linkage.

Data modification: Altering data in some way so as to control disclosure. Data modification techniques include: data swapping, noise addition, and rounding. Sometimes called perturbative or data distortion methods.

Data reduction: Disclosure control methods that work by restricting the data rather than distorting it. Examples are sampling, variable deletion and aggregation/recoding. Sometimes also called non-perturbative methods, metadata-level controls or non-perturbative masking.

Data release: Any process of data dissemination where the data custodian 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, called population units, (whether persons, households, businesses, or other entities). Normally used to mean microdata (i.e. not summary/aggregate statistics).

Data share: A dynamic data situation where the data custodian has made a decision to allow a fixed set of entities access to a given dataset.

Data situation: The relationship between data and its environment.

Data subject: Given data that is personal information, and therefore about a person, that person is the data subject.

Data unit: A case in a dataset; a set of data about a single population unit or data subject.

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.

De-identified information:

- In this book, information that is not (or is no longer) about an identified or reasonably identifiable individual, and does not reveal personal information about such an individual.
- In the Privacy Act: `personal information is de-identified if the information is no longer about an identifiable individual or any individual who is reasonably identifiable'.

De-identification: A process involving the removal or replacing of direct identifiers in a dataset, followed by the application of any additional techniques or controls required to remove, obscure, aggregate, alter and/or protect data in some way so that it is no longer about an identifiable or reasonably identifiable individual. This will usually require that the risk of other types of disclosure, such as attribute disclosure or inferential disclosure, are also very low.

Differential privacy: A privacy standard or model popular in the computer science academic literature. The principle underlying differential privacy is that the presence or absence of any single individual record in a data set should be unnoticeable when looking at the results of analysis on the dataset.

Disclosure:

- In common use: the inappropriate association of information to an individual, via re-identification or attribute disclosure.
- In the strict legal context of the Privacy Act: when one entity makes data (containing personal information) accessible or visible to others outside the entity, and the subsequent handling of the personal information is released from the entity's effective control. The release may be a proactive release, a release in response to a specific request, an accidental release or an unauthorised release by an employee.

Disclosure control methods: Methods used to reduce the risk of disclosure. They are usually based on reducing the amount of, or modifying, the data to be released.

Disclosure risk: This is expressed as the probability of a disclosure.

Direct identifier: A variable that can be used to uniquely identify an individual, either alone or together with other direct identifiers, and often in combination with other readily available information.

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.

Functional de-identification: A holistic approach to de-identification which asserts that data can only be determined as de-identified or not in relation to its environment.

Harmonisation: The process of recoding a variable on a dataset so that it more directly corresponds to an equivalent variable on another dataset.

Indirect identifier: A variable that can be used to identify an individual with a high probability, either alone or together with other indirect identifiers, and in combination with auxiliary

information. Almost any variable can be an indirect identifier, depending on the auxiliary information available to the intruder.

Inferential disclosure: An inferential disclosure occurs if the dissemination of a dataset enables the intruder to obtain a better estimate for a confidential piece of information than would be possible without the data.

Intruder: A data user who attempts to disclose information about a data subject through identification and/or attribute 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*.

k-anonymity: A privacy standard that requires at least k records within a dataset that have the same combination of indirect identifiers. Common thresholds are k = 3 or 5.

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. In the context of disclosure risk assessment and reduction, key variables are normally indirect identifiers common to the target dataset and the auxiliary information. An intruder launching a linkage attack would compare the values of the key variables in the target dataset and the auxiliary information, since any matches could lead to re-identification.

License agreement: A permit, issued under certain conditions which enables a researcher to use 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 information.

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.

Open data: Data released without any access restrictions, usually by publishing on the internet.

Penetration test: A component of disclosure risk assessment involving replicating what a plausible motivated intruder might do (and the auxiliary information and resources they might have) to execute a re-identification and/or disclosure attack on some data. Also known as intruder test.

Personal information: A term defined in Section 6(1) of the Privacy Act, which 'means any information or an opinion about an identified individual, or an individual who is reasonably identifiable:

(a) whether the information or opinion is true or not; and

(b) whether the information or opinion is recorded in a material form or not'.

Under the Privacy Act, this term can only refer to living individuals.

Population: The set of units from which a dataset is drawn. 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, whether a person, household, business or other entity.

Privacy: A concept that is much discussed and debated, but for which there is no unequivocal definition. While the concept of privacy is a very broad one, the Privacy Act relates primarily to information privacy. Information privacy 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 their beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

Reasonably identifiable: An individual will be reasonably identifiable for the purposes of the definition of personal information in s 6(1) of the Privacy Act where

(a) it is technically possible for re-identification to occur (whether from the information itself, or in combination with other information that may be available); and

(b) there is a reasonable likelihood that this might occur.

Re-identification: The discovery of the identity of individual(s) in an apparently de-identified dataset.

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:

1. access to a whole range of raw (protected) data and process it themselves or
2. access to outputs, e.g. tables from the data.

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 laboratory whereby access to a dataset can be controlled.

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 sampling 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.

Scenario analysis: A framework for analysing plausible data intrusion attempts. This framework identifies (some of) the likely factors, conditions and mechanisms for disclosure,

including establishing the key variables that might be used by a data intruder to re-identify data units.

Sensitive information: A defined category of personal information under the Privacy Act, this includes information or opinion about a person's racial or ethnic origin, political opinion, religious or philosophical beliefs, sexual orientation, criminal record and health, genetic and/or biometric information. This information is accorded a higher standard of protection under the APPs. For example, an entity requires a person's consent before they can collect sensitive information about them.

Sensitive variables: Distinguishable from sensitive information, which is a legal term, 'sensitive variables' are variables contained in a data record that the data subjects would not want to be disclosed. Sensitive variables are subjective and cannot be exhaustively defined, however they would include sensitive information as described above, and any other type of personal information that a data subject wants to keep confidential. For example, this could include data related to income, wealth, credit record and financial dealings.

Spontaneous recognition: This occurs where an individual is sufficiently unusual in a data collection, or the data user knows a sufficient number of an individual's attributes such that the user might make the unintentional observation they have identified the individual within the dataset.

Statistical Disclosure Control (SDC): An umbrella term for the integrated processes of disclosure risk assessment, disclosure risk management and data utility.

Synthetic data: Data that have been generated from one or more population models, designed to be non-disclosive.

Target: Object of interest to an intruder, and thereby subject to attack. Applies to an individual, a record, a variable, some information or a dataset.

Acknowledgements

The authors gratefully acknowledge the invaluable contributions and input from:

- Sarah Ghali (Office of the Australian Information Commissioner)
- John Newman (Australian Bureau of Statistics)

We also thank the following for their insightful and useful comments on drafts:

- Melanie Drayton (Office of the Australian Information Commissioner)
- Stephen Hardy (Data61, CSIRO)
- Arthur Street (Data61, CSIRO)
- Australian Bureau of Statistics
- Australian Institute of Health and Welfare
- Australian Government Department of Employment

Part of the development work for the original UK version of this publication, the 'Anonymisation Decision-Making Framework' was conducted through a series of one-day meetings of members of the United Kingdom Anonymisation Network (UKAN) core network between 2012 and 2014 and we 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 the original UK version 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 (CSIRO, Australia)
- Malcolm Oswald (HSISC, UK)
- Natalie Shlomo (University of Manchester, UK)
- Jeni Tennison (Open Data Institute, UK)
- Mathew Woollard (University of Essex, UK)

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].

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

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. & DALE, A. (1999) Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk; Netherlands Official Statistics, Spring 1999: 6-10, available 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., O'HARA, K. & TUDOR, C. (2016a) The Anonymisation Decision-Making Framework. UKAN Publications.

ELLIOT, M. J., MACKEY, E., O'SHEA S., TUDOR, C. & SPICER, K. (2016b) 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. & 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.

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].

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.

ITNEWS (2016a) Health pulls Medicare dataset after breach of doctor details. Available at https://www.itnews.com.au/news/health-pulls-medicare-dataset-after-breach-of-doctor-details-438463 Accessed 29 Apr 2017.

ITNEWS (2016b) Govt pulls dataset that jeopardised 96,000 employees. Available at https://www.itnews.com.au/news/govt-pulls-dataset-that-jeopardised-96000-employees-438809

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.

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

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.

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 OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2013) APP Guidelines, available at www.oaic.gov.au.

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2014a) Data breach notification – A guide to handling personal information security breaches, available at www.oaic.gov.au.

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2014b) Guide to undertaking privacy impact assessments, available at www.oaic.gov.au

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2014c) Privacy business resource 4: De-identification of data and information, available at www.oaic.gov.au

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2015) Guide to privacy regulatory action, available at http://www.oaic.gov.au

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2017a) Guide - What is personal information? Available at http://www.oaic.gov.au

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2017b) Australian Community Attitudes to Privacy Survey 2017, available at http://www.oaic.gov.au

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].

PRODUCTIVITY COMMISSION (2017) Data Availability and Use- Draft Report. Available at http:// http://www.pc.gov.au/inquiries/completed/data-access#report

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].

RITCHIE, F. Secure access to confidential microdata: four years of the virtual microdata laboratory. The Labour gazette, 2(5):29-34, 2008.

RITCHIE, F. & GREEN, E. (2016) Department of social services data access project final report. Project Report. Australian Department of Social Services, Canberra. Available from: http://eprints.uwe.ac.uk/31874

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.

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].

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].

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].

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].

CONTACT US

t 1300 363 400
+61 3 9545 2176
e csiroenquiries@csiro.au
w www.data61.csiro.au

FOR FURTHER INFORMATION

Dr Christine M O'Keefe PhD MBA
t +61 2 6216 7021
e Christine.O'Keefe@csiro.au
w www.data61.csiro.au

AT CSIRO WE SHAPE THE FUTURE

We do this by using science and technology to solve real issues. Our research makes a difference to industry, people and the planet.