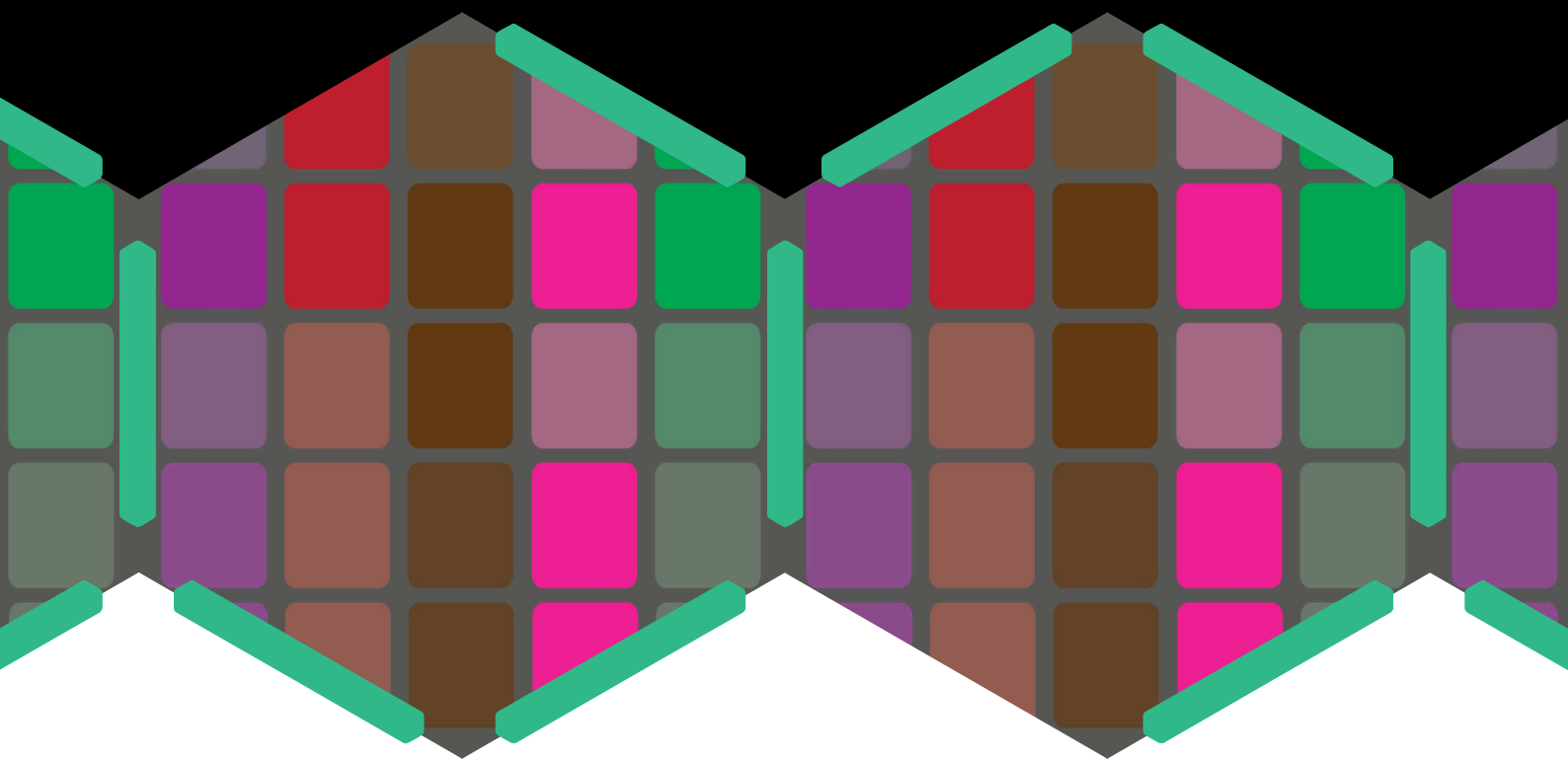
The De-Identification Decision-Making Framework

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

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

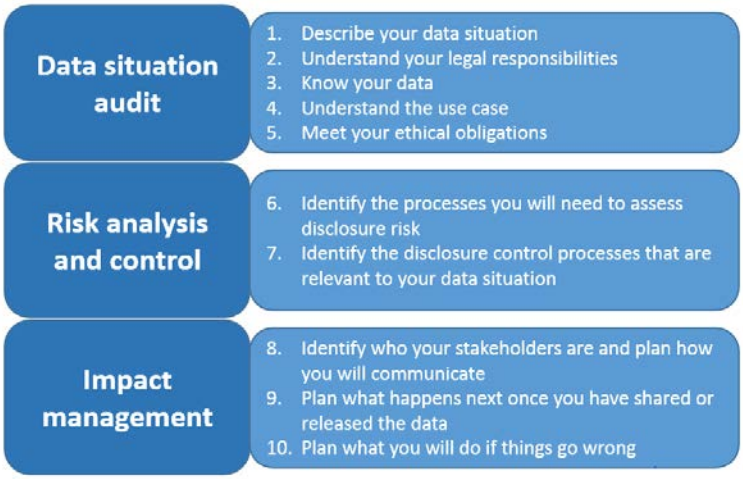
4. Secure virtual access: on approval, data is accessed via a secure link

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.

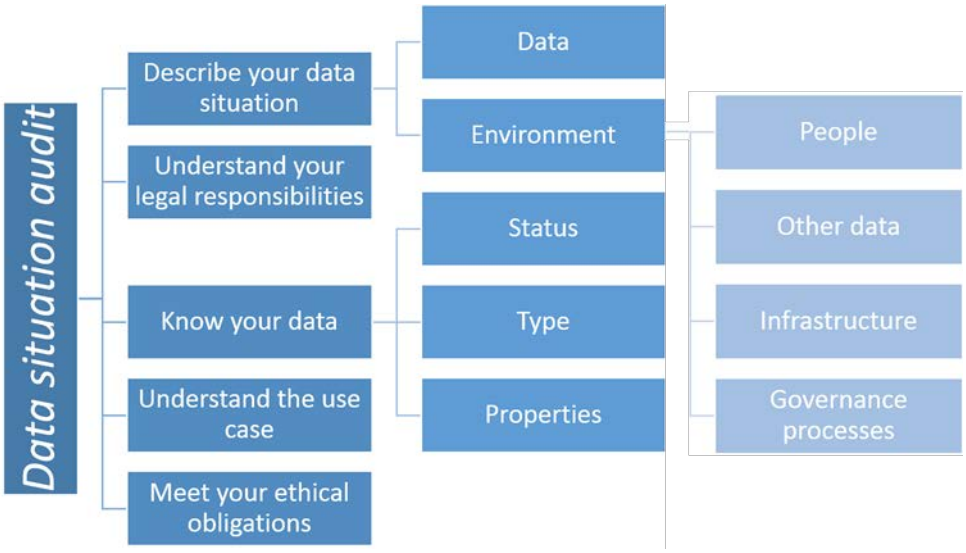
Risk analysis and control (Components 6-7) requires you to consider the technical processes that you will need to use in order to both assess and manage the disclosure risk associated with your data situation.

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.

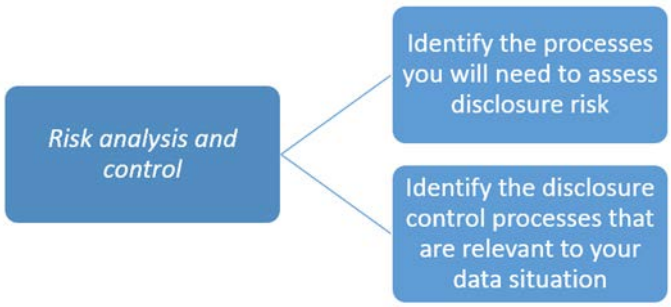
Component 5: Meet your ethical obligations

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

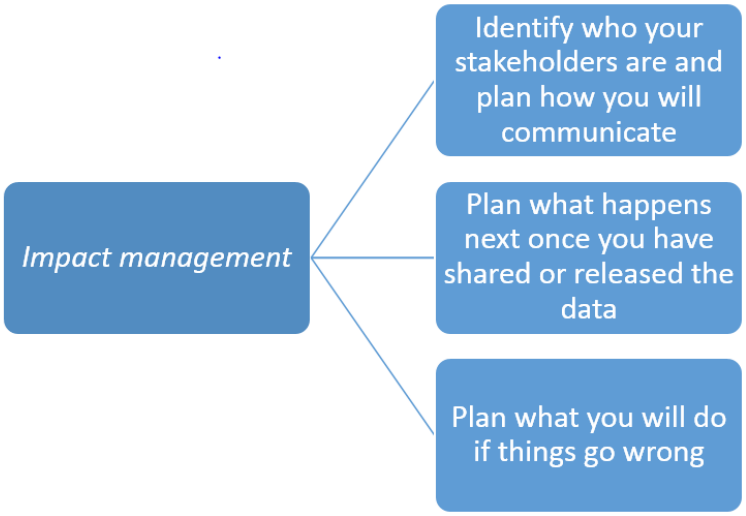
Disclosure control processes essentially attend to either or both of the two elements of your data situation: the data and its environment. If your risk analysis in Component 6 suggests that you need stronger controls then you have two (non- exclusive) choices:

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.

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.