ANDS Guide

De-identification

Level: Awareness

Last updated: 4 April 2018

Web link: http://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data

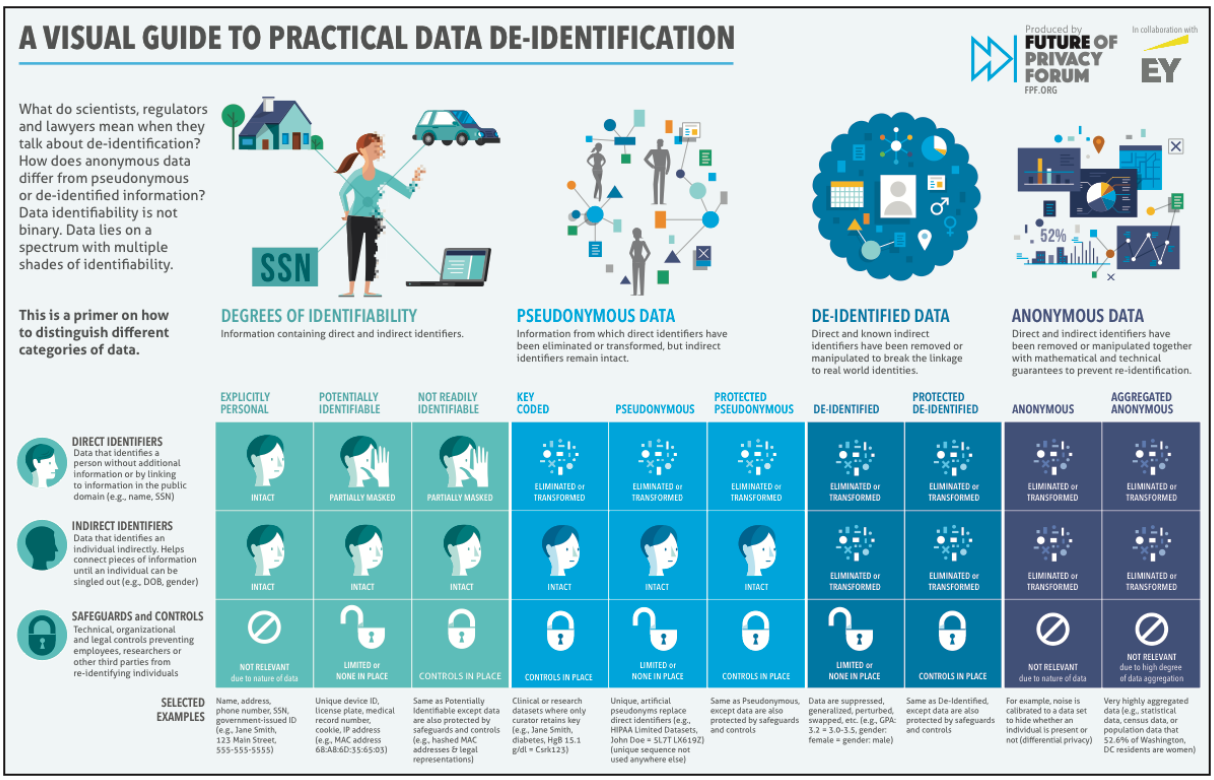
This Guide is intended for those who own a dataset and want to de-identify it for the purpose of sharing or publishing the data.

Introduction

De-identification is most commonly undertaken to protect the privacy of individuals. It aims to allow data to be used by others without the possibility of individuals being identified. Data de-identification may also be used to protect organisations, such as businesses or other information such as the spatial location of mineral or archaeological findings or endangered species.

De-identification may not be required, for example, in oral histories where it is customary to publish and share the names of people interviewed and for which they have given their consent.

The Future of Privacy Forum's visual guide to practical data de-identification (CC-BY-ND 4.0)



Definitions

In Australia, in addition to the Commonwealth legislation, almost each state and territory has its own privacy legislation. The Office of the Australian Information Commissioner offers links to all this legislation.

The Commonwealth *Privacy Act 1988* (Part II, Division I, Section 6) defines:

> **personal information** as "information or an opinion, whether true or not, and whether recorded in a material form or not, about an identified individual, or an individual who is reasonably identifiable." Common examples are an individual's name, address, telephone number, date of birth, bank account details and commentary or opinion about a person.
>
> **identification information** about an individual as: the individual's full name, alias, date of birth, sex; current, last known or previous address or employer, or driver's license number.
>
> **identifier** of an individual as a number, letter or symbol, or a combination thereof, that is used to identify or verify the identity of the individual, but does not include the individual's name. For instance, an identifier of an individual may include a Medicare number or Hospital/Medical Record Number.
>
> **de-identified** as "personal information is de-identified if the information is no longer about an identifiable individual or an individual who is reasonably identifiable". De-identified information is no longer considered personal information under the *Privacy Act 1988* and can be shared.

Anonymisation and confidentialisation are sometimes used interchangeably with de-identification.

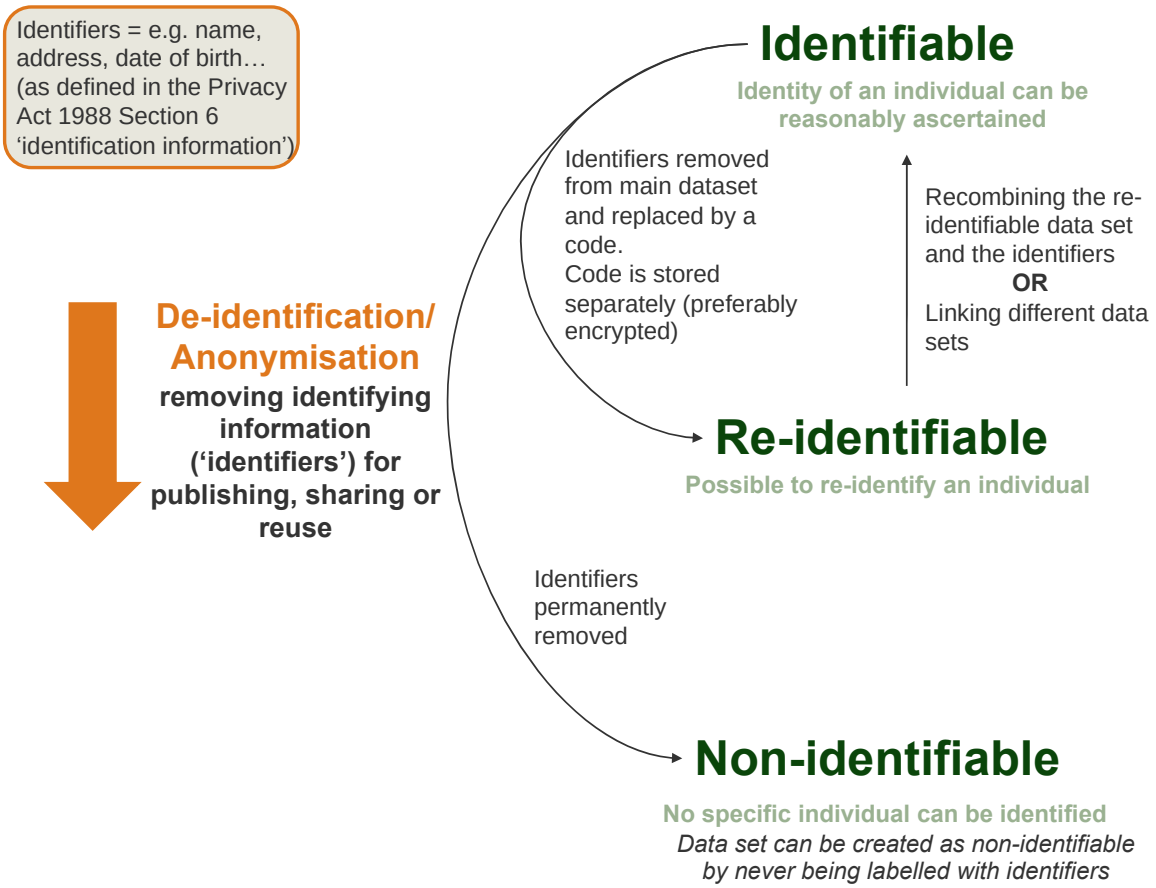
- **De-identification** is the removal of identifying information from a dataset, and this data could potentially be re-identified e.g. if the identifying information is kept (as a key) and recombined with the de-identified dataset.
- **Anonymisation** is the permanent removal of identifying information, with no retention of the identifying information separately.
- **Confidentialisation** is a less commonly used term, the National Statistical Service uses it to mean a process that involves both de-identifying data and then taking the additional step of assessing and managing the risk of indirect identification occurring in the de-identified dataset

`Identifiable, re-identifiable, non-identifiable' data

The National Statement on Ethical Conduct in Human Research (2007, updated May 2015), published by the National Health and Medical Research Council has three mutually exclusive definitions of data identifiability:

- Individually identifiable data
- Re-identifiable data
- Non-identifiable data

The National Statement avoids the term 'de-identified data', as it says it is ambiguous and could refer to both re-identifiable and non-identifiable data.



Australian practical guidance for de-identification

- The Australian Government's Office of the Australian Information Commissioner (OAIC) and CSIRO Data61 have a 'De-identification Decision Making Framework', which is a "practical guide to de-identification, focussing on operational advice".
- The OAIC also provides high-level guidance on de-identification of data and information, outlining what de-identification is, and how it can be achieved.
- The Australian Government's guidelines for the disclosure of health information, includes techniques for making a data set non-identifiable and example case studies.
- Australian National Statistical Service's information on confidentiality and how to confidentialise data
- Australian Bureau of Statistics' National Statistical Service Handbook. Chapter 11 contains a summary of methods to maintain privacy.
- Australian Bureau of Statistics' A good practice guide to sharing your data with others
- med.data.edu.au gives information about anonymisation
- Office of the Information Commissioner Queensland's guidance on de-identification techniques

International practical guidance for de-identification:

- USA government's guidance for methods for de-identifications of health information
- USA National Institute of Standards and Technology has two guides to de-identification
- The UK Anonymisation Network has a comprehensive (171 pages) Anonymisation Decision-Making Framework
- The UK Data Service's guide about anonymisation
- UK Research Data Network hosts a curated list of resources for managing personal data and best practice for anonymisation and preservation.
- The UK Information Commissioner's Office's Anonymisation: managing data protection risk code of practice

Qualitative data

- The UK Data Archive gives advice on anonymising qualitative data
- The Irish Qualitative Data Archive has developed a tool for anonymising qualitative data

Audio-visual data

Digital manipulation of audio and image files can be used to remove identifying information. However, techniques such as voice alteration and image blurring are labour-intensive and expensive and are likely to damage the research potential of the data. If confidentiality of audio-visual data is an issue, it is better to obtain the participant's consent to use and share the data unaltered, with additional access controls if necessary.

Tips for managing de-identification:

- Plan de-identification early in the research as part of your data management planning
- Retain original unedited versions of data for use within the research team and for preservation
- Create a de-identification log of all replacements, aggregations or removals made
- Store the log separately from the de-identified data files
- Identify replacements in text in a meaningful way, e.g. in transcribed interviews indicate replaced text with [brackets] or use XML markup tags e.g. <anon>.....</anon>

Management of identifiable data

Data may often need to be identifiable (i.e. contains personal information) during the process of research, e.g. for analysis. If data is identifiable then ethical and privacy requirements can be met through access control and data security. This may take the form of:

- Control of access through physical or digital means (e.g. passwords)
- Encryption of data, particularly if it is being moved between locations
- Ensuring data is not stored in an identifiable and unencrypted format when on easily lost items such as USB keys, laptops and external hard drives.
- Taking reasonable actions to prevent the inadvertent disclosure, release or loss of sensitive personal information.

Resources

- The Office of the Australian Information Commissioner has a very comprehensive guide to securing personal information.
- The med.data.edu.au discussion paper explores the legal, best practice and security frameworks in Australia for managing and storing sensitive data.
- The Research Data Network (UK) maintains a curated list of resources that covers management of personal data.

Five Safes: Working with identified data

The UK Data Service has developed the **Five safes framework** to provide secure access to carry out work that would not usually be possible with de-identified data:

- Watch the video (4 minutes) about the Five safes
- Read about the Five safes

It should be noted that de-identification is not a 'magical bullet' for being able to share and publish sensitive data. It should be considered within a range of activities to protect the privacy of research participants, such as obtaining informed consent for data sharing and controlling access to the data.

Related ANDS Guides

- Publishing and sharing sensitive data
- Data sharing considerations for Human Research Ethics Committees

Feedback?

We welcome your feedback on this guide. Please email contact@ands.org.au with any comments or questions.

About ANDS

The Australian National Data Service (ANDS) makes Australia's research data assets more valuable for researchers, research institutions and the nation.

ANDS is a partnership led by Monash University in collaboration with the Australian National University (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). It is funded by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to reuse and republish this work, or any part of it, with attribution to the Australian National Data Service (ANDS).