Human Services Dataset De-Identification and Five Safes Framework Policy

Summary: This policy outlines how de-identification and the Five Safes Framework are embedded into management of the Human Services Dataset

Document approval

The Human Services Dataset De-Identification and Five Safes Framework Policy has been endorsed and approved by:

Jessica Stewart

Executive Director, Family and Community Services Insights, Analysis and Research (FACSIAR)

Approved: 1 July 2021

Document version control

Distribution:	Public
Document name:	Human Services Dataset De-Identification and Five Safes Framework Policy
Trim Reference	D20/971530
Version:	Version 1.1 – July 2021 – Changes following governance transition to FACSIAR
This document replaces	Version 1.0 – June 2020 – Document created
Document status:	Revised
Authoring unit:	FACSIAR
Date:	1 July 2021
Next Review Date:	12 months from version date, or earlier if there are legislative or significant operational changes

Table of contents

1 Purpose of policy

1.1 Purpose

This policy is designed to:

- outline how the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Office of the Australian Information Commissioner (OAIC)'s De-Identification Decision-Making Framework (DIDMF) and NSW Government's Data Sharing Principles (based on the Five Safes Framework developed by the Office for National Statistics in the United Kingdom) have been incorporated in managing disclosure risk;
- reduce disclosure risk when the Human Services Dataset (HSDS) is released to approved data users for research and analysis for an approved purpose (Tier Two Data), or made publicly available (Tier Three Data);
- re-affirm the FACS Insights Analysis and Research (FACSIAR) HSDS Governance and Privacy Project Team's commitment to protecting the privacy of individuals and applying leading data privacy and security practices.

1.2 Background and policy links

The HSDS has been created by de-identifying and linking unit-level administrative records collected through NSW government agencies. Although de-identification significantly reduces disclosure risks when data is linked and used, it is possible that individuals can still be reasonably identifiable from the unit-record data, particularly if the individual exhibits unique characteristics or is part of a small population sample. In this regard, de-identification is '*an exercise in risk management, rather than an exact science*'[1].

To manage disclosure risk and to help the Project Team and government data partners make effective decisions about safe management of data, this policy provides guidance on application of leading industry standards. This policy refers to the following guidelines:

- CSIRO and OAIC, *The De-Identification Decision-Making Framework*;
- ABS, *'Managing the risk of disclosure: The five safes framework'*;
- Office of the National Data Commissioner, *Best Practice Guide to Applying Data Sharing Principles*.

[1] 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. Accessed on 23 June 2020.

The following documents are linked to this policy, and are available on the HSDS website:

- Human Services Dataset Data Governance Framework;
- Guidelines for access to and use of the Human Services Dataset;
- Human Services Dataset Release and Publication Policy.

2 Scope and application

This policy applies to data collected, stored and used for the HSDS.

This policy is to be followed by the Project Team, Data Linkage Centres, NSW Data Analytics Entity, and Approved Analysts.

3 Legislation

This policy supports compliance with the following legislation and policies:

- Public Interest Direction made under section 41(1) of the *Privacy and Personal Information Protection Act 1998* (NSW)
- Public Interest Direction made under section 62(1) of the *Health Records and Information Privacy Act 2002* (NSW)
- *Privacy and Personal Information Protection Act 1998* (NSW)
- *Health Records and Information Privacy Act 2002* (NSW)
- *Data Sharing (Government Sector) Act 2015* (NSW)
- NSW Cyber Security Policy
- NSW Data and Information Custodianship Policy
- NSW Government Information Classification, Handling and Labelling Guidelines.

4 Policy statement

The Project Team respects the privacy of individuals and takes privacy risks seriously. The Project Team recognises that proper de-identification is critical to promoting trust and meeting community expectations regarding re-uses of government administrative data for policy making, research and program design and evaluation. In addition to complying with the requirements of the Public Interest Direction and Health Public Interest Direction (PIDs), the Project Team applies the Five Safes Framework to data governance practices for the HSDS.

4.1 De-identification

Data de-identification is the first step in preparing agency datasets (Tier One Data) for linkage and to enable the use of the Tier Two Data for research and analysis. The PIDs mandate the use of the following de-identification techniques and processes.

4.1.1 Anonymisation

Tier One Data disclosed to the Centre for Health Record Linkage (CHeReL) includes Identifier Information and Service Usage Data. Identifier Information includes name, date of birth, address and gender. During the anonymisation process, CHeReL removes Identifier Information and allocates an arbitrary project specific person number (PPN) to each individual.

PPNs are generated according to the following requirements:

- it must be unique to the HSDS and the individual;
- it does not follow a pattern;
- it is not associated with any Identifier Information;
- it is not generated from any aspect of Personal or Health Information.

PPNs change from each successive version of the HSDS.

PPNs help prevent re-identification of individuals or a class of individuals and significantly reduces the possibility that any Personal or Health Information will be disclosed or released.

4.1.2 Privacy verification checks

Privacy verification checks apply to the de-identified, linked unit-level data (Tier Two Data). As detailed below, Approved Analysts must conduct these checks prior to commencing data analytics, and before the external or public release of any statistical outputs.

De-identification checks prior to data analytics

Prior to using Tier Two Data for analytics, the Data Analytics Entity and Approved Analysts must check that no Personal or Health Information has been included in the data.

Approved Analysts must not use the data for analysis until it has been cleansed of Personal or Health Information.

Aggregation

Prior to the external or public release of statistical outputs, Approved Analysts must ensure that the data is aggregated and no longer contains unit-level data. During this process, the Data Analytics Entity and Approved Analysts must check that no Personal or Health Information has been included in the statistical outputs.

Confidentialisation

Approved Analysts must confidentialise statistical outputs where an individual displays uncommon or unique characteristics that may cause them to be indirectly or directly identified, or is otherwise at a higher risk of re-identification. Confidentialisation is conducted in accordance with the Australian Bureau of Statistics' Confidentiality Information Series.

To be an approved data user, analysts and researchers must have sufficient technical skills to identify disclosure risks in the data provided for analysis, and in any statistical outputs they generate. Users must have an understanding of confidentialisation techniques to mitigate the risk of re-identification.

4.2 Five Safes Framework

As discussed above, a key risk of sharing the HSDS is the inclusion of Personal and Health Information in Tier Two and Tier Three Data, and the risk that individuals may be re-identified. The Project Team applies the Five Safes Framework in all practices, procedure and systems associated with the Project to ensure comprehensive management of disclosure risk.

Under the Five Safes Framework, data sharing risks are managed across five dimensions:

- Safe Project: use the data for appropriate and authorised purposes;
- Safe People: personnel have knowledge, skills and incentives to store and use the data appropriately;
- Safe Settings: ensure there are practical controls on the way the data is accessed and use the data in a safe and secure environment;
- Safe Data: apply appropriate protections to the data to minimise disclosure risk
- Safe Output: ensure that the identity of the individuals remain confidential in any external or public release of outputs.

The table below shows how the Five Safes Framework supports the data sharing controls applied across the three tiers of data of the HSDS.

Five Safes Framework	Tier One Data	Tier Two Data	Tier Three Data
Safe project	**High control over appropriateness of data use.** Tier One Data is used only for the purposes of carrying out data linkage services in accordance with the PIDs	**High control over appropriateness of data use.** Data is used for an approved project that is consistent with the Approved Purposes, with a valid analytics aim, for public benefit, and not for compliance or regulatory purposes (such as targeting a specific individual). This is underpinned by multiple layers of oversight and clear lines of accountability.	**There is no control required over appropriateness of data use**. Published Tier Three Data can be used for any purposes.
Safe people	**High access and use control.** Tier One Data can be accessed only by authorised personnel within the Data Linkage Centre which has authority to collect and use raw agency data (Tier One Data) for the purpose of linkage. Access is provided on a need-to-know basis and in accordance with the separation principle to ensure identifiable information is handled separately from service usage data. The Data Linkage Centre must comply with the NSW Cyber Security Policy and are obliged to comply with the PPIP Act and HRIP Act to the extent modified by the PIDs. Currently the Centre of Health Record Linkage (CHeReL) has been approved to undertake data linkage services for the HSDS.	**High access and use control.** The Data Analytics Entity must comply with the NSW Cyber Security Policy, the PPIP Act and HRIP Act to the extent modified by the PIDs. Approved Analysts are appropriately authorised to access and use the data. Approved Analysts will have: • undergone background checks (WWCC and National Police Check) • technical ability in data analysis • signed a legally binding Data Privacy and Confidentiality Agreement setting out the responsibilities and limits of data use • undertaken training in privacy and conditions of data use.	**There is no access and use control required**. Data can be accessed by researchers, universities, the Project Team, Participating Agencies and Public Sector Agencies. Tier Three Data can also be made publicly available, in which case anyone can access the data.

Safe settings	**High level of controls to prevent unauthorised use.** Tier One Data is accessed in a closed and secure linkage environment. In performing data linkage CHeReL complies with the NSW Cyber Security Policy. In particular, CHeReL has an ISO 27001 aligned Information Security Management System in place to govern data and IT assets. Secure File Transfer Protocol is used to transfer data to the Data Analytics Entity and Approved Analysts. The physical environment is also secure.	**High level of controls to prevent unauthorised use.** Tier Two Data is hosted securely with strict access security controls. The analytics work takes place on a secure analytics platform, accessible only to Approved Analysts and with appropriate monitoring (such as active supervision and the use of audit logs). The physical environment is also secure.	**There is no control required to prevent unauthorised use.** Tier Three Data is released externally to researchers and analysts who perform the analysis. It can also be made publicly available.
Safe data	**Appropriate controls to minimise disclosure risk.** Tier One Data contains Personal and Health Information which is subject to protection in accordance with the privacy legislation. Access to Tier One Data is provided on a need-to-know basis and in accordance with the separation principle to ensure identifiable information is handled separately from service usage data. Identifier information are removed from Tier One Data and replaced with arbitrary PPNs.	**Appropriate controls to minimise disclosure risk.** Personal identifiers are removed from project data and controls are in place to prevent misuse and re-identification. This involves the Data Analytics Entity and Approved Analysts checking Tier Two Data in a restricted and secure environment for identifying information, before it is used for analytics work.	**High level of controls to minimise disclosure risk.** Tier Three Data is de-identified, aggregated and, where needed, confidentialised to protect privacy of individuals.

Safe outputs	**Not applicable** No outputs are produced using Tier One Data. Identifier information are removed from Tier One Data and replaced with arbitrary PPNs. The resultant dataset becomes Tier Two Data.	**High level of controls to prevent disclosure risk** Outputs are produced at an aggregate and de-identified level and do not contain unit-level data. The Data Analytics Entity, Approved Analysts and Project Team are required to check outputs for disclosure risk, including inadvertent disclosure, before outputs are externally released or made public. Steps could include: • aggregating outputs to a minimum cell size of ten • suppressing the cell so that the output at risk is not disclosed • expanding the output to include additional data, therefore increasing the sample size • considering any additional contextual information that would place the output at greater risk of disclosure, and applying disclosure control techniques as needed (such as perturbation, small cell suppression or omitting the data) • adopting the ABS's Statistical Area Levels (for location-based statistics) or Confidentialisation (where a high disclosure risk exists due to uncommon or unique traits).	**High level of controls to prevent disclosure risk** Released Tier Three Data is de-identified and aggregated data, which does not include any Personal and Health Information.

5 Roles and responsibilities

The main roles and responsibilities for the implementation of this policy are as follows:

5.1 Data Custodian

- The Deputy Secretary of Strategy, Policy and Commissioning in DCJ serves as the Chair of the Stronger Communities Data Partnership (SCDP) and has overriding custodianship, control and responsibility for data that enters the HSDS. In May 2020 the Secretary of the Department of Communities and Justice delegated responsibility as Data Custodian of the HSDS to the Deputy Secretary Strategy, Policy and Commissioning. Consequently, the Deputy Secretary is the Data Custodian, responsible for ensuring compliance with the PIDs and the appropriateness of the security systems and processes in place to protect data.
- The Data Custodian notifies the NSW Privacy Commissioner of re-identification (deliberate or inadvertent) of de-identified data, other than in accordance with the privacy verification checks.

5.2 The Centre for Health Record Linkage

- The Centre for Health Record Linkage (CHeReL) as the NSW Data Linkage Centre removes identifying information from Tier One Data and assigns PPNs to create Tier Two Data.

5.3 Data Analytics Entity and Approved Analysts

- Apply confidentialisation where an individual displays uncommon or unique characteristics, or is otherwise at higher risk of re-identification.
- Conduct privacy verification checks (Information Protection Gates) before any statistical outputs are externally released to ensure that information is sufficiently de-identified and no longer falls within the definition of Personal and Health Information.
- Notify the Project Team of any deliberate or inadvertent re-identification of de-identified data, other than in accordance with the privacy verification checks.

5.4 FACSIAR HSDS Governance and Privacy Team

- FACSIAR as the Project Team, implements this policy and ensure that all relevant personnel associated with the Project are aware of policy requirements
- The Project team also reviews outputs prior to release and consults with the Human Services Data Governance Advisory Committee on interpretation risks.

6 Monitoring, evaluation and review

It is the responsibility of the FACSIAR Data Privacy & Governance Team to monitor and update this policy when required. This policy will be reviewed every year and when any significant new information, legislative or organisational change warrants amendments to this document.

7 Support and advice

You can get advice and support about this policy from the FACSIAR HSDS Governance and Privacy Team who has carriage of this document: dataprivacy@facs.nsw.gov.au

If you are reviewing a printed version of this document, please refer to our website to confirm that you are reviewing the most recent version of the policy. Following any subsequent reviews and approval this policy will be uploaded to the Intranet and all previous versions removed.

8 Definitions

The table below is a list of terms, keywords and/or abbreviations used throughout this document.

Term	Definition
Analytical Services	The study, analysis, modelling, research or evaluation of Tier Two Data.
Approved Analysts	The Project Team or a person (including a researcher or analyst) that: a) has been approved by the Data Custodian to provide Analytical Services for and on behalf of the Project Team; and b) is under a contractual obligation to comply with the PPIP Act and HRIP Act to the extent modified by a relevant Public Interest Direction.
Approved Purpose	Any activity, task, work, step, process or measure that facilitates or enables the Project Objectives.
De-identified information	De-identified information is information from which the identifiers about the person have been permanently removed, or where the identifiers have never been included. This means that the information is not personal information for the purposes of the PPIP Act. *Source: https://www.ipc.nsw.gov.au/fact-sheet-de-identification-*

Term	Definition
	personal-information
FACSIAR	Family and Community Services Insights, Analysis and Research
Five Safes Framework	An internationally recognised risk management model that is designed to help identify and manage data sharing risks. Under this framework, data sharing risks are managed across five 'safety' dimensions: people, projects, settings, output and data. *Source: https://data.nsw.gov.au/data-sharing-principles*
Health Information	The meaning of Health Information is given in section 6 of the HRIP Act. For the purpose of this policy, Health Information includes information set out in Part A of Schedule 1 of the *Public Interest Direction under section 62(1) of the HRIP Act*.
HRIP Act	*Health Records and Information Privacy Act 2002* (NSW)
Human Services Data	Data or information (which may include Personal Information and Health Information) within a Participating Agency's or its contractors' or agent's records or system in connection with a Public Sector Agency's or other government agencies' interaction with, or provision of supports, services or programs to, an individual.
Human Services Dataset	The Human Services Dataset (HSDS) is a de-identified, longitudinal dataset that integrates administrative data collected on children and young people born after 1 January 1990, and includes their parents, carers and guardians. The data consists of service streams, outcomes and life events collected from across NSW government agencies and covers: • Child protection; • Housing; • Justice; • Health; • Mental health; • Alcohol and other drugs (AOD); • Parental and perinatal risk indicators.
Information Protection Gates	The privacy verification process and checks that will be undertaken by the Data Analytics Entity or an Approved Analyst in accordance with the Public Interest Directions and before information held by the Data Analytics Entity or an Approved Analyst is externally released or disclosed to:

Term	Definition
	a) ensure compliance with the Public Interest Directions; b) ensure that only de-identified information is released or disclosed to a third party; and c) prevent re-identification of information by a third party, including a Participating Agency.
Personal Information	The meaning of Personal Information is given in section 4 of the PPIP Act. Personal Information means information or an opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is apparent or can reasonably be ascertained from the information or opinion. For the purpose of this policy, Personal Information includes the information set out in Part A of Schedule 1 of the Public Interest Direction made under *section 41(1) of the PPIP Act*.
PPIP Act	*Privacy and Personal Information Protection Act 1998* (NSW)
Project	Their Futures Matter project as described in the Public Interest Directions.
Public Interest Direction	Public Interest Direction made under section 41(1) of the PPIP Act or Public Interest Direction made under section 62(1) of the HRIP Act. Public Interest Directions are made by the NSW Privacy Commissioner to waive or make changes to the requirements for a public sector agency to comply with an Information Protection Principle or a Health Privacy Principle.
Tier One Data	Human Services Data that has not been through any de-identification process to remove any Personal Information and Health Information.
Tier Two Data	Data derived from Tier One Data that has Identifier Information removed and been allocated a PPN in accordance with a relevant PID.
Tier Three Data	Aggregated Tier Two Data that has been through the Information Protection Gates process in accordance with a relevant PID.