Government of South Australia

Privacy Committee
Of South Australia

Privacy and Open Data Guideline

Guideline

Version **1**

Executive Officer
Privacy Committee of South Australia
c/o State Records of South Australia
GPO Box 2343
ADELAIDE SA 5001
Phone (08) 8204 8786
privacy@sa.gov.au

Table of Contents

Introduction

This guideline aims to assist agencies to understand and address the risks to privacy when considering the public release of government datasets through the Government's Declaration of Open Data policy. The Guideline has been developed to ensure compliance with the South Australian Government's Information Privacy Principles Instruction (IPPI).

The South Australian Government is committed to government data being 'open by default' and has directed that agencies should release government data proactively and that it be published in accessible formats and available online.

In making its data open by default, the Government must also maintain high standards of privacy in the data it releases. The definition of Open Data means non-personal corporate data. Personal information of private citizens will not be released through Open Data.

Examples of information to be released under the Open Data program include a table of government spending on infrastructure projects or a dataset consisting of geocodes for public facilities. However, other agency data intended for release may not have such a clear cut distinction between non personal and personal information and may include de-identified personal information. De-identification of personal information is the removal of obscure personal identifiers and personal information so that identification of individuals, that are the subject of the information, is no longer possible.

Information privacy in the South Australian Government

Information Privacy in the South Australian Government is guided by the IPPI. The IPPI is an Instruction of Cabinet that is issued as *Premier and Cabinet Circular No 12*. It is the responsibility of the Principal Officer of a public sector agency to ensure that their agency complies with the IPPI. The IPPI sets out ten Information Privacy Principles (IPPs) that guide the way South Australian government agencies collect, store, use and disclose personal information. These Guidelines should be read in conjunction with the IPPI.

Under the IPPI, the term 'personal information' means:

> *Information or an opinion, whether true or not, relating to a natural person or the affairs of a natural person whose identity is apparent, or can reasonably be ascertained, from the information or opinion.*

Personal information is, therefore, any information that can be linked to an identifiable living person. This definition of personal information includes sensitive information. It could include information detailing the person's name, address, date of birth, financial or health status, ethnicity, gender, religion, alleged behaviour, licensing details, or a combination of such details. The important question to ask in determining whether information is personal information is whether it can identify a particular individual.

For the purposes of the IPPI a natural person is taken to be a living person. It does not extend to the information of the deceased. However, agencies should consider very carefully the status of the information of the deceased. In some cases there may be other legal restrictions on the publication of information of identified deceased individuals (eg a confidentiality or secrecy provision in a relevant Act).

Given the IPPI only applies to personal information, and not to data that has been de-identified care must be taken to determine that the information is properly de-identified and is not reasonably able to be re-identified.

Information Security Management Framework

Privacy classification is an important component of the Government's Information Security Management Framework (ISMF). Agencies should ensure that their information systems maintain standards of information security proportionate to the sensitivity of the information; this includes ensuring appropriate classification of information. Agencies are required to comply with the ISMF and any information security procedures developed by their agency.

It is recommended that, once a privacy risk assessment and mitigation techniques are undertaken, an appropriate ISMF classification is applied. Agencies can seek further advice regarding information security from their Agency Security Adviser or their IT Security Adviser.

The privacy risks of open data

While there are significant economic, democratic and social benefits to the release of government data, it can pose risks to the privacy of personal information. The primary risk to privacy in the release of government data is the identification of individuals. That is releasing data that is personal information or can be made into personal information through easily linking with other information.

The harms of identification of an individual in the release of a government dataset can be significant. A variety of harms could be reasonably anticipated from such identification, including:

- cause humiliation, embarrassment or anxiety for the individual, for example from a release of health data, it might be concluded that an individual accessed treatment for a sensitive sexual health condition
- impact on the employment or relationships of individuals
- affect decisions made about an individual or their ability to access services, such as their ability to obtain insurance
- result in financial loss or detriment
- pose a risk to safety, such as identifying a victim of violence or a witness to a crime.

The nature and extent of harm would depend on the type of data released and the extent of any identification of individuals. There are two key types of identification risks associated with the release of government data: spontaneous recognition of an individual and deliberation recognition.

Spontaneous recognition is the risk that identification is made without any deliberate attempt to identify a person. This can result from the release of a dataset that includes the data of individuals with rare characteristics. The risk of identification is proportionate to the rarity of the characteristic.

Deliberate recognition is the risk associated with a malicious or deliberate attempt to identify a person from the released dataset. This can result from list matching, or matching common characteristics in the released dataset to other publicly available datasets or information. It can also result from targeting a particular individual using a characteristic in the dataset already known by the person attempting to identify them.

Assessing the risks of identification of individuals in the release of government data is one of the necessary steps an agency must take to mitigate those risks to an acceptable level when making a decision whether to release data.

Managing the risks of identification in open datasets

Assessing the risks

The first step to managing privacy risks in the release of a public sector dataset is to undertake an initial assessment of the risk of making that data publicly accessible. Assessing this risk will require a detailed consideration of the data to be released. Methods used to assess this risk include:

- determining any specific unique identifying variables, such as name.
- cross-tabulation of other variables to determine unique combinations that may enable a person to be identified, such as a combination of age, income, postcode.
- Acquiring knowledge of other publicly available datasets and information that could be used for list matching.

The level of privacy risk will be dependent on the likelihood that identification could occur from the release of the data and the consequences of such a release. The level of risk will determine what steps the agency takes to mitigate the privacy risks.

The likelihood of identification

The most obvious factor to consider in the likelihood of identification is the presence of obvious identifying variables in the data, such as a name, date of birth or street address. Even with the absence of such variables the following factors need to be considered:

Motivation to attempt identification – Consider whether an individual or organisation would receive any tangible benefit (malicious or otherwise) from identification of individuals in the dataset.

Level of detail disclosed by the data – The more detail included, the more likely identification becomes. Where the dataset contains multiple variables for the same record-subject, identification could be made through the combination of those variables.

Presence of rare characteristics – If there are rare or remarkable characteristics for a record-subject the chances of identification are increased. For example, a 19 year old girl who is widowed is likely to be noticeable in the data.

Presence of other information – Even if the dataset itself does not include any data that would identify an individual it may include variables that can be matched with other information or datasets to identify a person.

Determining the potential consequences of identification

It is important to think about the potential consequences that might arise from the identification of individuals within a dataset. This includes the harm that might be suffered by those individuals identified and the impacts on the agency and government of such a data breach.

Once the level of risk has been determined the agency can consider the appropriate approach to mitigate that risk.

Mitigating risks through de-identification

A number of techniques can be applied to properly de-identify the dataset and mitigate any risks of identification of an individual. Consideration should be given to obtaining the proper approvals to undertake de-identification within the agency.

Removing identifiers

The most basic method of de-identification is to remove obvious identifying variables from the data such as an individual's name or address. For example, consider the following data:

Name	Address	Postcode	Age	Gender	Profession	Annual Salary
Barry Johns	10 Smith Street Woodville SA	5011	52	Male	Driving Instructor	$75,000

By removing basic identifiers this can become:

Postcode	Age	Gender	Profession	Annual Salary
5011	52	Male	Driving Instructor	$75,000

While on the face of it this data has been stripped of its identifiers, it retains a relatively high potential for re-identification: the data still exists on an individual level and other, potentially identifying, information has been retained. For example, some South Australian postcodes have very small populations and combining this data with other publicly available information, can make re-identification a relatively easy task.

While it may be tempting for agencies to strip out all potentially identifying information, doing so could render the data meaningless. The fact that somewhere in Australia there is a driving instructor that earns $75,000 may have limited potential use.

Pseudonymisation

A related method of de-identification is 'pseudonymisation' which involves consistently replacing recognisable identifiers with artificially generated identifiers, such as a coded reference or pseudonym. In the example above, Barry Johns would be assigned a randomly selected numerical value:

Individual reference	Postcode	Age	Gender	Profession	Annual Salary
SR23597	5011	52	Male	Driving Instructor	$75 000

Pseudonymisation allows for different information about an individual, often in different datasets to be correlated without the consequence of direct identification of the individual. For example, the information above could be correlated with:

Individual reference	Marital status	Number of children	Highest level of education attained	Number of cars owned by household
SR23597	Divorced	2	Diploma	3

However, pseudonymisation also has a relatively high potential for re-identification, as the data exists on an individual level with other potentially identifying information being retained. Also, because pseudonymisation is generally used when an individual is tracked over more than one dataset, if re-identification does occur more personal information will be revealed concerning the individual

Reducing the precision of the data

Rendering personally identifiable information less precise can make the possibility of re-identification more remote. Dates of birth or ages can be replaced by age groups; specific salaries can be replaced by salary ranges.

For example Barry John's data now becomes:

Name	Postcode	Age range	Gender	Profession	Annual Salary range
SR23597	5011	50-60	Male	Driving Instructor	$60,000 - $80,000

Related techniques include suppression of cells with low values or conducting statistical analysis to determine whether particular values can be correlated to individuals. In such cases it may be necessary to apply the frequency rule by setting a threshold for the minimum number of units contributing to any cell. Common threshold values are 3, 5 and 10.

For example, applying a threshold value of 3 to the following table the cell indicating the number of driving instructors at ages 35-40 has a value less than 3 may be suppressed or aggregated into a bigger range.

Age	Postcode	Number of Driving Instructors	Average Annual Salary
25-30	5011	20	$50,000
35-40	5011	**2**	$60,000
45-50	5011	10	$65,000

More advanced techniques include introducing random values or 'adding noise'. It may also include altering the underlying data in a small way so that original values cannot be known with certainty but the aggregate results are unaffected.

Aggregation

Individual data can be combined to provide information about groups or populations. The larger the group and the less specific the data is about them, the less potential there will be for identifying an individual within the group. An example of aggregated data would be:

Initial data:

Profession	State	Annual Salary	Number of drivers
Driving Instructor	South Australia	$49,500	200
		$40,000	10,000
		$45,000	2,000
		$56,000	3,748
		$58,000	11,414
		$66,000	31,203

Aggregated data:

Profession	State	Annual Salary	Number of drivers
Driving Instructor	South Australia	<$50,000	12,200
		>$55,000	46,365

Tools to assist in de-identification

A range of tools and software packages are available to assist in the task of de-identifying datasets. These tools provide an automated method of applying a particular de-identification

method and may assist an agency to determine with more precision the success of the de-identification method applied and the privacy risk of public release of the dataset.

Some examples of these tools are listed below. The Privacy Committee does not endorse the use of any particular tool and provides these examples for information only. Agencies should conduct their own research to determine any tools suited to their de-identification task. It is noted that these tools have been developed in other jurisdictions and, therefore, some of their functionality may not be applicable to the Australian data environment.

Mu-ARGUS – Statistics Netherlands

Privacy Analytics Risk Assessment Tool

SUDA – University of Manchester

Cornell Anonymisation Toolbox

University of Texas Anonymisation Toolbox

Testing de-identification and reassessing the risk

It is good privacy practice to test the methods that the agency has employed to mitigate the privacy risks of publishing the dataset. Primarily this will involve attempting to re-identify individuals from the de-identified dataset. This type of testing is sometimes referred to as penetration testing.

In testing the de-identification method by attempting to re-identify the dataset, consideration should be given to all the factors considered in the initial assessment of identification risk, including the:

- presence of unique or clearly identifying variables
- presence of rare characteristics
- cross tabulation of variables to identify unique or rare combinations of variables for the same data subject
- availability of other data that could be linked with the dataset and lead to identification.

The test should meet the following criteria:

- The test should attempt to identify particular individuals and one or more private attributes relating to those individuals.
- The test may employ any method which is reasonably likely to be used by a motivated intruder, that is, a person motivated to find out identified information.
- The test should use any lawfully obtainable data source which is reasonably likely to be used to identify particular individuals in the datasets.

The agency should consider whether it is necessary to engage any specialist knowledge or expertise to properly test its methods of de-identification. Testing would need to be conducted by trusted parties in secure environments to avoid any inadvertent disclosure of personal information.

The agency should reassess the risk of identification once it has tested the vulnerability of its dataset. If the risk is now at an acceptable level and a person could not be identified from

the dataset then the agency can publish the dataset. If the risk of identification remains high the agency would have to consider whether it can employ further methods of de-identification to mitigate this risk. If the risk of identification is not able to be mitigated to an acceptable level, the dataset should not be released.

See Appendix 1 for a privacy risk flow chart.

SA NT DataLink

It may not be possible to eliminate all the risks and agencies must consider if the data is suitable to release through an Open Data scheme. Another way of allowing secure access to the information is through the use of on-site data laboratories. One such data laboratory is SA NT DataLink. The Privacy Committee recommends that agencies only use data laboratories where that data is available at the custodian level or, where this is not possible, the privacy and ongoing security needs of the information is assured. Agencies should undertake a Privacy Impact Assessment where this is the case.

SA NT DataLink is an unincorporated joint venture comprising South Australian and Northern Territory Government and non-government organisations. It enables the linkage of administrative and clinical datasets to allow population level health, social, education and economic research and evidence-based policy development to be undertaken with de-identified data, minimising risks to individual privacy when compared to traditional sample based research using identified data.

Data linkage through SA NT DataLink is supported by the Privacy Committee through the granting of a number of exemptions. The exemptions allow State Government agencies to disclose limited identifying variables, such as name, date of birth and address, to SA NT DataLink for inclusion in its Master Linkage File to enable the creation of links between multiple government datasets. The exemptions are subject to strict conditions on the governance of data.

For further information about SA NT DataLink please see https://www.santdatalink.org.au/animation. Alternatively, SA NT DataLink may be contacted on telephone 8302 1604 or email to santdatalink@unisa.edu.au.

Where do I get more information?

This Guideline has been issued by the Privacy Committee of South Australia. The Committee exists to:

- advise on measures that should be taken to protect personal information
- refer written complaints received about breaches of privacy to the relevant authority
- consider agency requests for exemption from compliance with the IPPI.

Further information about the IPPI is available at www.archives.sa.gov.au/privacy.

Other relevant documents

Short Guide to the Information Privacy Principles

Acknowledgements

In developing this guidance the Privacy Committee has utilised the significant work of other Australian and international privacy authorities on the issue of online privacy. This includes the Office of the Queensland Information Commissioner, Dataset Publication and De-identification Techniques.

Appendix 1 – privacy risk flow chart