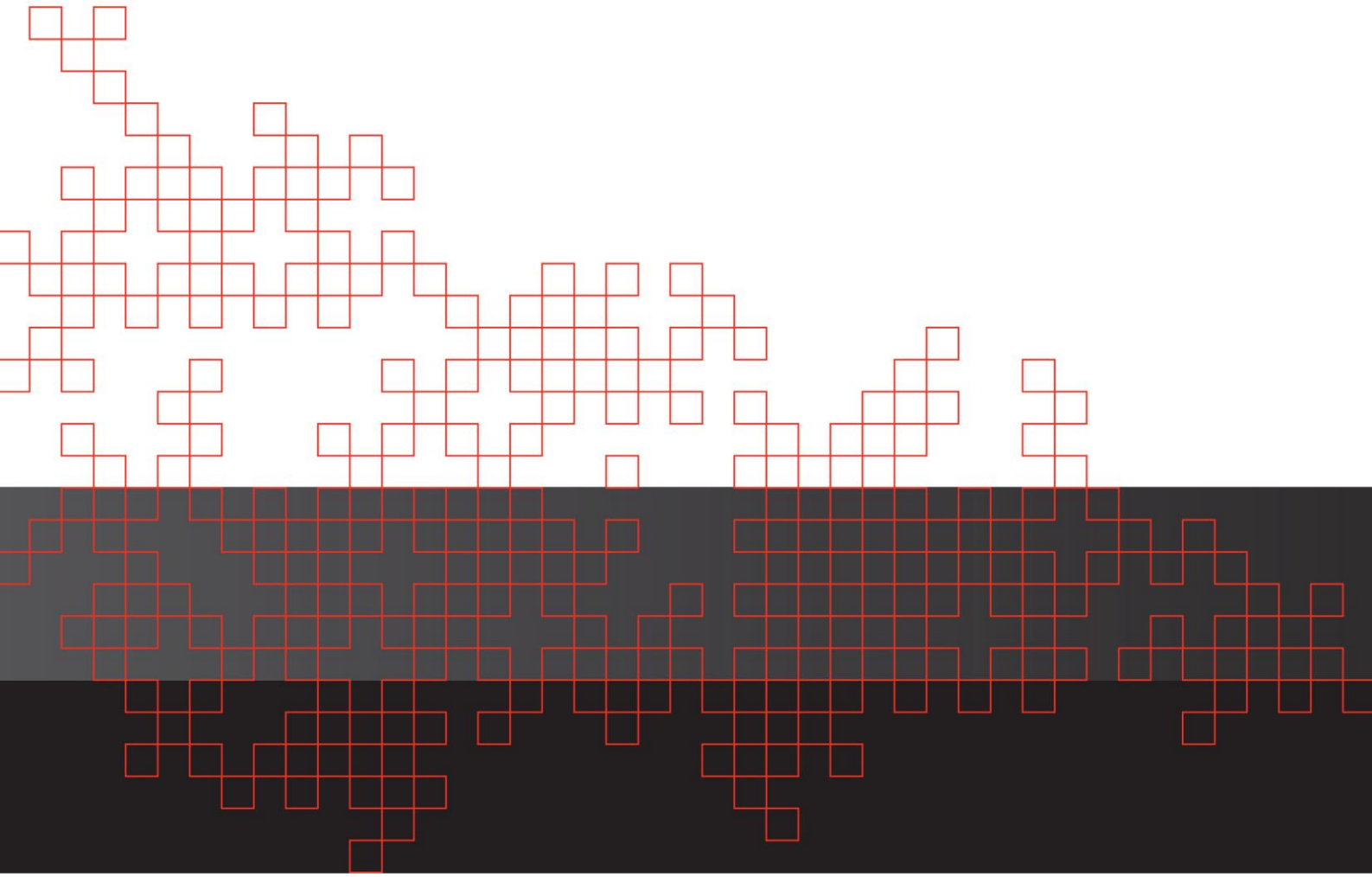
Data confidentiality policy and procedures

CRIME STATISTICS AGENCY



Contents

1. Introduction

The Crime Statistics Agency's (CSA) statistical outputs present aggregate data about incidents that come to the attention of Victoria Police and are entered onto the Law Enforcement Assistance Program (LEAP). The LEAP database includes information about criminal and non-criminal incidents (such as family incidents) recorded by police, as well as information about individuals and how they are involved in these incidents, namely victims and/or offenders.

The CSA has an obligation to protect the privacy of individuals and has implemented clear policies to ensure people are unlikely to be identified through data released by the agency.

2. Document purpose

This document outlines the agency's data confidentiality policy, the principles applied in confidentialising the data, and the processes implemented to ensure data confidentiality.

3. Scope

This policy is applied to all statistical outputs released by the CSA. Exemptions in the application of this policy may only be made by the Chief Statistician, with reference to applicable aspects of the *Crime Statistics Act 2014*, Victorian Protective Data Security Standards and the broader Victorian Privacy Framework.

Other security and data management policies and practices that the agency apply in order to maintain the security of the law enforcement data held within its databases are outside the scope of this document.

4. Definition of key terms

For the purposes of this policy, the following statistical terms and definitions from the OECD Glossary of Statistical Terms and Australia's National Statistical Service are used:

- *statistical disclosure control* – "the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released."[i]
- *data confidentiality* – "data confidentiality is a property of data, usually resulting from legislative measures, which prevents it from unauthorized disclosure" and "...refers to the procedures in place to prevent disclosure of confidential data, including rules applying to staff, aggregation rules when disseminating data, provision of unit records, etc."[ii]
- *confidentialisation* – "confidentialisation 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."[iii]
- *peturbation based disclosure control methods* – methods applied at the dissemination stage, which change the data prior to release which adequately reduce risk of disclosure and maintain data confidentiality, while ensuring the core "information content" is preserved.[iv] In other words, the risk of identification is lowered but the story that a reader of the output can obtain from the data is unchanged.

The cognate legal and privacy term is:

- *de-identification* – the process of removing or altering information, after which that information is not relatable to a reasonably identifiable person. This usually involves two steps – the removal of direct personal identifiers such as name and address, and removing unique or rare characteristic(s) of an individual which would be sufficiently novel to potentially enable the person's identification.[v] The definition of this privacy term is similar in meaning to the statistical term *confidentialisation*.

5. Why is confidentiality necessary in statistical data?

Statistical disclosure control processes are vital in meeting legal and ethical obligations to protect the identity and privacy of individuals and organisations, while still ensuring the usefulness of the data for statistical and research purposes. Statistical disclosure control methodologies are applied by statistical organisations around the world, using a variety of different methodologies and working to different acceptable standards of disclosure risk, depending on applicable legislative frameworks.

While confidentialisation can be used to manage the risk of identification, it may not completely eliminate risk. Judgements are required about whether presentation of certain data are still likely to reveal private information about individuals, and whether it should be released publicly.

6. Confidentiality, data security and privacy

The CSA manages its law enforcement data in accordance with Victorian Protective Data Security Standards, and with reference to the *Crime Statistics Act 2014*. Under these standards, all unit record data is stored securely within the CSA, and transformed to create crime statistics. Data are only released after this transformation process has been completed, and data has been aggregated and confidentiality processes have been applied appropriately[1]. By default, the CSA confidentialises all aggregate data displaying small person-based counts and small counts of sensitive offence-based data.

7. De-identification and confidentialisation of data

There are two primary methods for protecting the privacy of an individual within a dataset which contains social or personal information; de-identification of data and confidentialising of data.

De-identification of data refers to the stripping of some or all personally identifying features from the raw, unit record data (e.g. name and address)[vi].

Confidentialisation refers to the process of removing or altering information or collapsing detail in aggregate data to reduce the risk of a person or organisation being identified in the data (either directly or indirectly).

[1] The only exception to this practice under the *Crime Statistics Act* 2014 is where the Chief Statistician is compelled to provide unit record law enforcement data by prevailing legislation.

There are two general methods (often referred to as statistical disclosure control methods) used to confidentialise data for publication:

1. data modification methods (also known as perturbation) which involve changing the data slightly to reduce the risk of disclosure, while retaining the overall integrity of the data content; and
2. data reduction methods which aim to control or limit the amount of detail available, without compromising the overall usefulness of the information.

In practice, de-identification and confidentialisation requires three key steps:

1. the removal of any direct identifiers (e.g. name and address) from the raw data;
2. aggregation of the data; and/or
3. assessment and management of the risk of indirect identification occurring in the de-identified and aggregated data, often through direct perturbation based disclosure control methods.

8. What types of recorded crime data are confidentialised and why?

8.1 Measures produced by the CSA

The CSA releases information based on five main population types:

1. recorded offences – This measure is a count of the number of criminal offences recorded in LEAP within a reference period.
2. alleged offender incidents – This measure is a count of the number of alleged offenders processed by police within a reference period. There may be multiple incidents within the reference period that involve the same individual, business or organisation as an offender.
3. victim reports – This measure is a count of the number of times an individual, business or organisation reports that they have been a victim of one or more criminal offences to Victoria Police and a record is subsequently made in LEAP.
4. family incidents – This measure is a count of incidents attended by Victoria Police where a Victoria Police Risk Assessment and Risk Management Report (also known as an L17 form) was completed. This population also includes information about Affected Family Members and Other Parties involved in family incidents.
5. unique person-level data – This measure is a count of unique victims and offenders that come into contact with police within the reference period. Regardless of the number of times within that reference period that they have been in contact with police, they will represent just one person in this data.

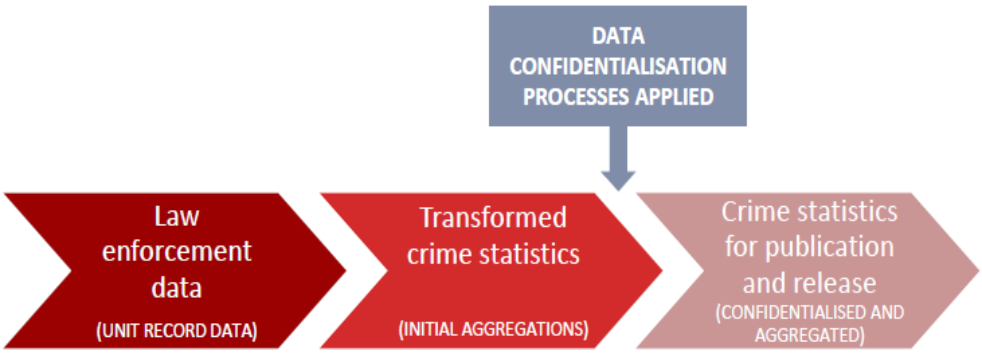
Data related to recorded offences is not attached to an individual victim or offender and therefore, in the majority of cases, this measure is not confidentialised. The exception to this guideline is where the data covers sensitive crime types (i.e. homicide and sexual offences), or where the data is focussed on a small geographic area.

The person-based measures (victim reports, alleged offender incidents, unique persons and affected family members and other parties involved in family incidents) are subject to confidentialisation to ensure the anonymity of individuals is protected, where numbers are small and there is a potential risk that a person may be identified from the data published.

9. At what stage are data confidentialisation processes applied?

Data confidentialisation is applied prior to formal release of statistical data by the CSA.

Figure 1: Application of key data confidentialisation processes in the transformation of data and preparation for dissemination.



10. Aggregating and confidentialising data to maintain individuals' privacy

10.1 Aggregating data to avoid confidentialisation

In order to provide meaningful statistics, every effort is made to avoid the need to confidentialise published data. One of the simplest methods used is data aggregation, as shown in the example below:

Table 1: Unconfidentialised data (small cells highlighted)

Age	Males	Females	Persons
10–14	54	30	84
15–19	77	70	147
20–24	88	80	168
25–29	80	74	154
30–34	67	60	127
35–39	45	35	80
40–44	44	49	93
45–49	30	33	63
50–54	25	20	45
55–59	16	14	30
60–64	4	5	9
65–69	1	2	3
70+	25	14	39
Total	556	486	1042

Table 2: Data confidentialised by aggregating row variables

Age	Males	Females	Persons
10–14	54	30	84
15–19	77	70	147
20–24	88	80	168
25–29	80	74	154
30–34	67	60	127
35–39	45	35	80
40–44	44	49	93
45–49	30	33	63
50–54	25	20	45
55–59	16	14	30
60–64	4	5	9
65+	26	16	42
Total	556	486	1042

In this example, the age grouping of 65-69 may be grouped with the 70+ category to avoid the use of confidentialisation methods. All other column headings would remain the same, and the 70+ would be relabelled to 65+.

Where a significant amount of other detail from the table may be lost as a result of aggregation, alternative methods can be used. There may also be instances where data is already aggregated to the highest practical level and small counts cannot be avoided.

In these cases, to maintain the confidentiality of individuals while still providing a detailed level of data, values with small cells (less than or equal to 3) can be confidentialised using perturbation based disclosure controls, which will hide the true number of these values, as outlined in section 10.2.

10.2 Confidentiality methodology

The CSA confidentialises cells that are between 1 and 3. This is denoted in the tables by the value "≤ 3" appearing in cells with small numbers. For the purpose of calculating row and column totals, each cell between 1 and 3 is assigned a value of 2, regardless of the true value of that cell. This methodology allows for totals to be calculated in tables with small cells, but does mean that totals for some variables may differ across tables within a publication or set of data cubes. In this way, the overall picture shown within the data can be preserved, but the true number is never known for certain.

The following example demonstrates the difference between the unconfidentialised data table (table 3) and a table where confidentialisation methods have been applied (table 4).

Table 3: Unconfidentialised data (small cells highlighted)

Age	Males	Females	Persons
10–14	54	30	84
15–19	77	70	147
20–24	88	80	168
25–29	80	74	154
30–34	67	60	127
35–39	45	35	80
40–44	44	49	93
45–49	30	33	63
50–54	25	20	45
55–59	16	14	30
60–64	4	5	9
65–69	1	2	3
70+	25	14	39
Total	556	486	1,042

Table 4: Confidentialised data

Age	Males	Females	Persons
10–14	54	30	84
15–19	77	70	147
20–24	88	80	168
25–29	80	74	154
30–34	67	60	127
35–39	45	35	80
40–44	44	49	93
45–49	30	33	63
50–54	25	20	45
55–59	16	14	30
60–64	4	5	9
65–69	≤ 3	≤ 3	4
70+	25	14	39
Total	557	486	1,043

This confidentialisation method is known as “frequency-level confidentialisation”, and is method used in all applicable outputs released by the CSA.

10.3 Differencing between tables and different data releases

In the context of statistical disclosure controls, 'differencing' refers to the possibility that multiple cuts of data generated from the same dataset, when confidentialised, may still enable the possible identification of very small (1 or 2) values by reverse calculating from the observed differences between values produced across confidentialised tables. Differencing is a challenge to conventional statistical confidentialisation techniques, and has been the subject of considerable deliberation throughout the international statistical community. While the CSA will take due care in the application of confidentiality procedures, the agency cannot completely control for all possibilities of differencing over time.

Exemptions to the CSA confidentiality policy

The default policy position of the CSA is to apply this confidentiality policy to all statistical outputs from the agency. Exemptions to this policy may only be made by the Chief Statistician, having regard to all relevant applicable legislative frameworks and the likely risk of any such release. The CSA may also release unconfidentialised data to another entity when compelled by prevailing legislation.

[i] Statistics Netherlands, Statistics Canada, Germany FSO, University of Manchester, 2005, Glossary of Statistical Disclosure Control, incorporated in paper presented at Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 October 2005. Reproduced in the OECD Glossary of Statistical terms. URL: http://stats.oecd.org/glossary/detail.asp?ID=6996. Retrieved 6 November 2014.

[ii] Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000. Reproduced in the OECD Glossary of Statistical terms. URL: http://stats.oecd.org/glossary/detail.asp?ID=4930. Retrieved 6 November 2014.

[iii] National Statistical Service, Confidentiality Information Series, National Statistical Service website, www.nss.gov.au/nss/home.NSF/pages/Confidentiality+Information+Sheets. Summarised in Office of the Australian Information Commissioner (2014) Privacy business resource 4 De-identification of data and information. April 2014. URL: http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf Retrieved: 6 November 2014.

[iv] Statistics Netherlands, Statistics Canada, Germany FSO, University of Manchester, 2005, Glossary of Statistical Disclosure Control, incorporated in paper presented at Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 October 2005. Reproduced in the OECD Glossary of Statistical terms. URL: http://stats.oecd.org/glossary/detail.asp?ID=6950. Retrieved 6 November 2014.

[v] Based upon the definition in the Commonwealth s 6(1) of Privacy Act.and reproduced in: Office of the Australian Information Commissioner (2014) Privacy business resource 4 De-identification of data and information. April 2014. URL: http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf Retrieved: 6 November 2014.

[vi] Australian Bureau of Statitsics (2014) SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing. February 2014. URL: http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/1f951c3977938593ca257ab500090a08/$FILE/SSF%20Guidance_Geographic%20Differencing_1.pdf Retrieved: 9 September 2014.