



Office of the Information Commissioner
Queensland

Privacy and Public Data

Managing re-identification risk



The Office of the Information Commissioner licence this report to the Queensland Legislative Assembly under a Creative Commons – Attribution License. People reading or using this report may do so under the following conditions: Attribution (BY), requiring attribution to the original author.

© The State of Queensland (Office of the Information Commissioner) 2020.

Copies of this report are available on our website at www.oic.qld.gov.au and further copies are available on request to:

Office of the Information Commissioner

Level 7, 133 Mary Street, Brisbane, Qld 4000

PO Box 10143, Adelaide Street, Brisbane, Qld 4000

Phone 07 3234 7373 or Freecall 1800 OIC QLD (1800 642 753)

Email administration@oic.qld.gov.au

Web www.oic.qld.gov.au

ISBN: 978-0-6484026-8-8

July 2020

The Honourable Curtis Pitt MP
Speaker of the Legislative Assembly
Parliament House
George Street
Brisbane QLD 4000

Dear Mister Speaker

I am pleased to present *Privacy and public data: Managing re-identification risk*. We prepared this report under section 135 of the *Information Privacy Act 2009*.

The report outlines how two Queensland government agencies manage privacy risks when releasing de-identified data. It makes recommendations to all government agencies.

In accordance with subsection 193(5) of the Act, I request that you arrange for the report to be tabled in the Legislative Assembly on the next sitting day.

Yours sincerely

A handwritten signature in black ink, appearing to read 'Rachael Rangihaeata'.

Rachael Rangihaeata
Information Commissioner

Table of contents

1. Summary and recommendations	1
Introduction	1
Conclusion	1
Key findings	2
Recommendations	3
2. Context	5
Privacy and public data	5
De-identification	6
Re-identification	7
Assessing re-identification risk	11
Treating re-identification risk	12
Audit objective	14
Audit scope	14
3. Releasing de-identified data	15
Introduction	15
Conclusion	15
Registers and custodians	16
Roles and responsibilities	17
Governing policies	18
Guidance for decision-makers	20
Monitor and review procedures	22
4. Managing re-identification risk	25
Introduction	25
Conclusion	26
Assessing re-identification risk	26
Analysing re-identification risk	28
Reviewing re-identification risk	34

1. Summary and recommendations

Introduction

Public data supports transparent and accountable government. The benefits of public data include evidence-based policy design, innovation, and better service delivery. All Queensland government agencies are encouraged to proactively release data on public platforms. This supports the 'push model' and the proactive disclosure aims of the *Right to Information Act 2009*.

While the majority of public data is not about people, this report is exclusively concerned with public datasets that contain, or are derived from, personal information. Agencies often 'de-identify' datasets containing this type of information prior to public release to meet their obligations under the *Information Privacy Act 2009*.

De-identification can be a useful tool that allows agencies to maximise the information they publish.¹

However, de-identified data is at risk of 're-identification'. When agencies release de-identified data on public platforms, they must adequately manage this risk to protect the identity of individuals and their personal information. Government acts as custodian of the personal information it holds on behalf of the community, who expect government agencies to safeguard their information.

This audit assessed whether two Queensland government agencies:

- have appropriate governance arrangements to manage the privacy risks of de-identified public data
- identify and manage privacy risks when releasing de-identified public data
- monitor and review the privacy risks of released de-identified data and update their mitigation strategies in response to environmental changes.

We have not named the audited agencies in this report. This is to protect the privacy of individuals with personal information in the examined datasets.

Conclusion

Agencies should manage privacy risks in public data the same way they manage risks in other activities. This includes identifying and assessing the risk of re-identification for

¹ In this audit, the term 'de-identified data' means data to which de-identification methods have been applied.

each dataset, applying the appropriate treatments to reduce the risk to an acceptable level, and monitoring and reviewing the risk periodically.

Sound governance arrangements support effective privacy risk management. They help decision-makers select fit-for-purpose de-identification techniques that balance data utility against the risk of re-identification.

Privacy risks are not static. They evolve in an environment where more information is continuously released, and new technologies emerge. Agencies that regularly review privacy risks and assess the effectiveness of risk treatments can better respond to environmental changes and manage risks appropriately over time. Documenting risk assessments and the reasons for selecting risk treatments helps regular monitoring and review.

Inadequate privacy risk management can lead to re-identification and the disclosure of personal information. When public data is re-identified, it can have serious consequences for stakeholders, clients and staff.

This audit highlights the importance of taking a methodical and robust risk management approach when releasing de-identified data on public platforms. Agencies that adopt good practices will be well placed to consider re-identification risk and protect individuals' privacy, including for vulnerable members of the community.

Key findings

Managing privacy and de-identified public data

Both audited agencies have detailed governance arrangements for public data in general. However, only one agency has adequate guidance to assist decision-makers when releasing de-identified data. The other agency's guidance is not sufficient to support effective re-identification risk management. As a result, its governance arrangements are not adequate to manage the privacy risks of de-identified data.

Neither agency has appropriate governance arrangements to regularly monitor and review re-identification risk in de-identified datasets. Without these arrangements, neither agency can be confident that risk management strategies remain effective over time.

Releasing and maintaining de-identified data

We examined four public de-identified datasets in each agency. Neither agency could consistently demonstrate how it developed de-identification techniques and managed re-identification risk in all four datasets.

One agency has sufficient records of re-identification risk management for two examined datasets. The other two datasets from this agency, and all four datasets from the other agency, lack sufficient records. We cannot assess how re-identification risk was managed in these datasets.

When assessing the re-identification risk in the published data, we assigned relatively low risk scores to datasets for one agency. This agency uses de-identification techniques to effectively reduce the risk of re-identification to generally low levels.

The other agency has significantly higher risk scores, noting three datasets scored medium or above in our assessment. There is a real risk of re-identification in these three datasets.

Neither agency monitors and reviews re-identification risk in the examined datasets. This means neither agency has assurance that the risk management strategies adopted for these datasets stay effective over time.

Recommendations

We made specific recommendations to each audited agency. The agencies accepted all recommendations.

The audit raised critical issues relevant to all Queensland government agencies. We make one recommendation to all government agencies², and four recommendations to all agencies that publish de-identified data.

² 'All Queensland government agencies' means all government agencies subject to the *Information Privacy Act 2009* (Qld) including Queensland government departments, statutory bodies, local governments, public universities, Hospitals and Health Services, and other public authorities.

We recommend all government agencies:

1. Review all published data and identify datasets containing de-identified data.

We recommend all government agencies that publish de-identified data:

2. Assign a custodian to each published de-identified dataset and capture this information in a register.
3. Implement and maintain policies or procedures that govern de-identified data releases, including guidance to decision-makers.
4. Monitor the external data environment and the effectiveness of risk treatments, and regularly review existing de-identified datasets for changes in re-identification risk.
5. Manage privacy when publishing de-identified data by adequately capturing, assessing and treating re-identification risk.

2. Context

Privacy and public data

Public data supports transparent and accountable government. The Queensland Government Open Data Policy Statement outlines the benefits of public data, including evidence-based policy design, innovation, and better service delivery.³ All Queensland government agencies are encouraged to proactively release data on public platforms consistent with obligations under the *Right to Information Act 2009* (Qld).

The majority of public data is not about individuals. For example, the Queensland Government Open Data Portal contains data about weather, public infrastructure, land development, and the location of government facilities.⁴

This report is exclusively concerned with datasets that contain, or are derived from, personal information. The *Information Privacy Act 2009* (Qld) defines personal information as:

information or an opinion, including information or an opinion forming part of a database, whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion (section 12)

Under the *Information Privacy Act 2009*, Information Privacy Principle 11 provides that personal information must not be disclosed to a third party, unless certain exceptions apply.⁵

Some public datasets contain information about an individual's gender, age, access to services and the location of individuals at a point in time. To meet their privacy obligations, agencies can 'de-identify' datasets containing this type of information prior to public release. This means they apply de-identification techniques to the data to remove or mask identifying information about the individuals. As a result, the published data is about people, but does not contain 'personal information'.

³ *Queensland Government Open Data Policy Statement* – viewed at <https://www.data.qld.gov.au/resources/documents/qld-data-policy-statement.pdf>

⁴ Viewed at - <https://www.data.qld.gov.au/>

⁵ For an overview of the Information Privacy Principles, refer to OIC guidance: <https://www.oic.qld.gov.au/guidelines/for-government/guidelines-privacy-principles/key-privacy-concepts/overview-of-the-information-privacy-principles>

De-identification can be a useful tool that allows agencies to maximise the information they publish. However, de-identification does not guarantee that privacy risks are managed, as de-identified data can be ‘re-identified’.

For example, in August 2019, the Office of the Victorian Information Commissioner issued a report on Public Transport Victoria’s disclosure of ‘myki’ travel information during a datathon.⁶ The report detailed how de-identified data releases can result in serious privacy breaches when appropriate controls are absent or ineffective.⁷

The myki report echoed similar findings by the Office of the Australian Information Commissioner, who reported in March 2018 that the Department of Health had breached Australian privacy principles when releasing de-identified data about Medicare and Pharmaceutical Benefits Schedule.⁸

In both instances, agencies had applied a range of de-identification techniques to the data to protect individuals. Despite de-identification, both examples experienced re-identification events.

Re-identification can reveal the identity of individuals and disclose their personal information. Unlike individual privacy breaches, re-identification events have the capacity to impact large groups of people.

Re-identification events can also undermine public confidence and trust in government agencies, discouraging others from releasing information. For these reasons, agencies must effectively manage privacy risks if releasing de-identified data.

De-identification

It is sometimes possible to apply techniques to data containing personal information to make it safe for public release. We call this process ‘de-identification’.

De-identification alters data to reduce the likelihood of disclosing personal information. Some de-identification techniques are simple, such as removing part of a dataset prior to publication.

⁶ OVIC, *Report of Investigation: Disclosure of myki Travel Information* – viewed at: https://ovic.vic.gov.au/wp-content/uploads/2019/08/Report-of-investigation_disclosure-of-myki-travel-information.pdf

⁷ For key lessons from OVIC’s myki report, see: <https://ovic.vic.gov.au/blog/myki-incident-lessons-for-organisations/>

⁸ OAIC, *MBS/PBS Data Publication* – viewed at: <https://www.oaic.gov.au/privacy/privacy-decisions/investigation-reports/mbspbs-data-publication/>

Other de-identification techniques are complex, such as differential privacy, which adds random noise to data while still allowing for accurate analytics on the dataset as a whole.⁹

The correct de-identification approach is context dependent. The appropriate technique depends on the format and sensitivity of each dataset.¹⁰ Agencies should also consider how de-identification might reduce data utility. In some cases, de-identified data may not be suitable for public platforms and should be shared through other means, such as under a data sharing agreement.

Regardless of the de-identification technique, agencies should never consider de-identification to be a fixed state. 'De-identified data' is simply data that has been through a de-identification process at one stage. It is not data where the risk of disclosing personal information has been permanently managed. This distinction is important, because while data may be thoroughly de-identified at a point in time, external events and new technology can quickly change this.

Recommendation 1

We recommend **all government agencies**:

Review all published data and identify datasets containing de-identified data

Re-identification

When individuals in de-identified datasets are identified, we call this 're-identification'. Re-identification events may breach the *Information Privacy Act 2009* and disclose personal information about individuals.

Re-identification often occurs when data is combined with auxiliary information to reveal information about an individual. Some examples of auxiliary information include:

- other public datasets and information, including social media

⁹ For guidance on de-identification techniques, refer to OIC guidance on *Privacy and De-identified Data*: <https://www.oic.qld.gov.au/guidelines/for-government/guidelines-privacy-principles/anonymity/privacy-and-de-identification>

¹⁰ For a detailed discussion on appropriate de-identification techniques, refer to *The De-identification Decision-Making Framework*, co-published by the OAIC and CSIRO: <https://data61.csiro.au/en/Our-Research/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework>

- non-public datasets, for example, a business' customer database
- personally observed information, for example, overhearing a conversation or witnessing an event.

While de-identified data may be safe in isolation, linking the data with auxiliary information can lead to re-identification.

For example, consider a fictional dataset about all people who received a government emergency disaster payment in 2019. This small dataset is de-identified, containing only high-level information about the payment recipients, as outlined in Figure 1.

Figure 1

Fictional dataset: All emergency disaster payments

Payment	Postcode	Recipient age	Disability
\$1000	4999	40-50	Yes
\$1000	4998	50-60	Yes
\$500	4998	50-60	No
\$1500	4997	85+	Yes

This dataset has two important characteristics. First, each row represents one person - we call this 'unit-level' information. This is common in public datasets. Second, there are 'unique' entries in the data. For example, there is only one person who received a \$500 payment, only one person who lives in postcode 4999, and only one person aged over 85 in this dataset.

Despite these unique entries, there is no information in the dataset about an individual whose identity is apparent or could be reasonably ascertained. While the dataset contains information about people (age, disability and location information), it is not possible to identify these people by looking at the data alone.

However, because of the unique unit-level entries, it is possible to combine the data with auxiliary information. Below are two simplified examples that illustrate one way re-identification could occur.¹¹

¹¹ For a comprehensive discussion on re-identification events, refer to *The De-identification Decision-Making Framework: Appendices*, co-published by the OAIC and CSIRO, page 18-22: <https://publications.csiro.au/rpr/download?pid=csiro:EP175702&dsid=DS1>

Observation event

A re-identification event may occur when a person obtains necessary auxiliary information through observation. At the extreme, observation events can involve malicious actors using social engineering to obtain information. But they can also occur when a person simply overhears a conversation or comes to know basic information about an individual.

For example, following a natural disaster, a neighbour living in postcode 4999 observes the house next door has been damaged. When talking with the owner of this house, the owner mentions they received an emergency disaster payment from the government.

If the neighbour accessed the example dataset in Figure 1, they could easily identify the homeowner as the first row in the data. This is because there is only one entry in the data for postcode 4999. Through this re-identification event, the data would reveal the homeowner received a \$1000 payment, is aged between 40 and 50, and has a disability.

Data linkage event

Data linkage events combine the original dataset with auxiliary data. These events can occur on a large scale and are often more complex than observational events.

For a simplified example, consider a non-government organisation that provides additional support to individuals who receive the emergency disaster payment discussed above. This organisation maintains records of the disaster victims it assists, as depicted in Figure 2.

Figure 2

Fictional dataset: additional relief to emergency payment recipients

Assistance	Recipient	Address
Emergency repairs	John Public	1 Example Street, Anytown QLD, 4999
Emergency repairs	Jane Citizen	1 Example Street, Othertown QLD, 4997

With this information, the organisation could quickly re-identify two entries in the original dataset, as outlined in Figure 3.

Figure 3

Fictional dataset: re-identified original dataset

Payment	Postcode	Recipient age	Disability
\$1000	4999	40-50	Yes
\$1000	4998	50-60	Yes
\$500	4998	50-60	No
\$1500	4997	85+	Yes

John Public (linked to 4999)

Jane Citizen (linked to 4997)

In this example, re-identification is possible because:

- The organisation knows John Public and Jane Citizen received the emergency disaster payment, and therefore their details must be in the original dataset.
- The organisation knows John Public lives in postcode 4999, and there is only one entry with this postcode in the original data.
- The organisation knows Jane Citizen lives in postcode 4997, and there is only one entry with this postcode in the original data.

Not only can the organisation re-identify two individuals in the original data, it can also learn new sensitive information.¹² While the organisation already knew John and Jane’s addresses, the data has revealed their age range, and that both have a disability.

Combinations of attributes

In our simplified example above, re-identification is possible by knowing information about single attribute (postcode). This is because the ‘postcode’ attribute contains unique values (4999 and 4997). Large public datasets are less likely to have attributes with unique values. However, re-identification can still be achieved when values across a combination of attributes are unique.

For example, in the revised dataset in Figure 4 below, no single attribute contains unique values. There are always two of each. Simply knowing somebody’s payment amount, postcode, age or disability status does not enable re-identification of individuals in the dataset.

¹² In this report, we use the term ‘sensitive information’ in a general sense. Note the *Information Privacy Act 2009* (Qld) contains a definition of sensitive information as it applies to the National Privacy Principles.

Figure 4

Fictional dataset: Emergency disaster payments without unique entries

Payment	Postcode	Recipient age	Disability
\$500	4999	85+	No
\$1000	4999	50-60	Yes
\$500	4997	50-60	No
\$1000	4997	85+	Yes

However, some entries in this dataset are unique when combining two attributes. For example, if a person overhears their co-worker talking about receiving a \$1000 disaster payment, and knows their co-worker is aged between 50 and 60, they can locate them as the second row in the dataset.

This is possible because only one entry, the second row in the data, is unique when combining the values '\$1000' and '50-60'. In this example, the data has revealed the co-worker lives in postcode 4999, and they have a disability.

Assessing re-identification risk

Agencies should already articulate their risk appetite and risk management principles in their existing risk frameworks.¹³ Enterprise risk principles equally apply when assessing and managing re-identification risk.

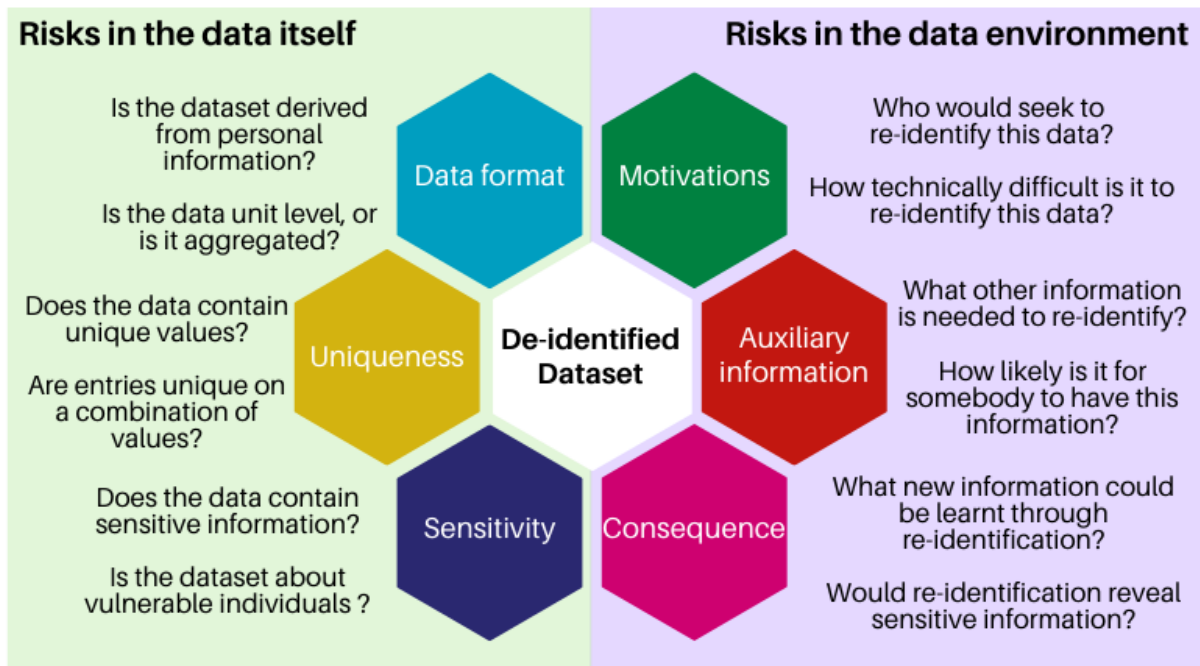
While it is not possible to foresee every possible re-identification scenario, agencies should conduct a detailed re-identification risk assessment prior to releasing de-identified data. Like any risk assessment, agencies must fully understand their public data activities and consider the wider risk environment.

It is not sufficient to simply ask 'how risky is the data?'. Agencies must also ask themselves 'how might a re-identification event occur?' and 'what new information could be learnt?'. Figure 5 outlines some high-level considerations when assessing re-identification risk.

¹³ For further detail on risk management frameworks, refer to Queensland Treasury, *A Guide to Risk Management*: <https://s3.treasury.qld.gov.au/files/guide-to-risk-management.pdf>

Figure 5

Re-identification risk assessment considerations



Source: Office of the Information Commissioner

When looking at the data itself, agencies should understand the data format, whether the data has unique values and if it contains sensitive information. When considering the wider data environment, agencies need to think about risk scenarios, including the 'who, what and how' of a re-identification event.

There are many ways to assess re-identification risk. Agencies may choose to assign qualitative risk descriptions or quantitative re-identification risk scores. No matter the approach, it is essential that agencies thoroughly consider and document all relevant risk factors.

Treating re-identification risk

There is no set threshold for acceptable re-identification risk. An agency's tolerance for re-identification risk should be informed by the type of data it releases, how important this data is for end users, and the potential impacts if the data is re-identified.¹⁴

¹⁴ For further detail on defining risk tolerance, refer to Queensland Treasury, *A Guide to Risk Management*, page 19: <https://s3.treasury.qld.gov.au/files/guide-to-risk-management.pdf>

While agencies should not release data where re-identification risk exceeds their agreed tolerance, it may be possible to reduce risk to an acceptable level. For example, if particular attributes in the data have unique values, these attributes can be further de-identified. Similarly, if the dataset contains sensitive information, the sensitive attributes can be removed to lower the risk of re-identification.

Agencies cannot control the wider data environment, including the motivations of data users and the volume and nature of available auxiliary information. Agencies also cannot control the secondary consequences of a re-identification event, should one occur.

Balancing privacy and data utility

Agencies face a privacy/utility trade-off when de-identifying data. It may be tempting to extensively de-identify data to lower re-identification risk, particularly where there are significant threats in the external environment. However, this can introduce new problems.

Effective de-identification may reduce data utility. In some cases, de-identification may render the data useless, or potentially misleading. Agencies must balance this trade-off when applying de-identification techniques to public data. Data that loses its utility through de-identification may not be suitable for publication at all.

Agencies should consider alternative release arrangements if environmental risk factors exceed their re-identification risk tolerance. For example, they can control access to the data and make it available only to trusted users, rather than releasing as public data.¹⁵

¹⁵ The Open Data Institute provides a helpful graphic of the 'data spectrum' which may assist agencies when considering appropriate data sharing methods: <https://theodi.org/about-the-odi/the-data-spectrum/>

Audit objective

The objective of this audit was to determine whether two Queensland government agencies adequately manage privacy risks when releasing de-identified data.

The audit assessed whether:

- Agencies have appropriate governance arrangements to manage the privacy risks of de-identified data.
- Agencies identify and address privacy risks when releasing de-identified data.
- Agencies routinely review the privacy risks of released de-identified data and update their mitigation strategies in response to environmental changes.

Audit scope

We examined policies and procedures, along with a selection of datasets. The audit focused on data released publicly, whether through an open data portal, a website or other access arrangements, such as hackathons.

The audit did not examine:

- datasets derived from raw data that contain no personal information, e.g. speed camera locations, bus routes
- non-public data sharing arrangements, e.g. data shared between government agencies, data shared with a third-party under a commercial contract.

Audited agencies

This audit examined two Queensland government agencies. We identified agencies through a risk assessment that considered the volume and sensitivity of released data. We have not named the audited agencies in this report. This is to protect the privacy of individuals with personal information in the examined datasets.

3. Releasing de-identified data

Introduction

Publishing data on public platforms is an effective way to proactively share information with the community. However, if agencies choose to publish de-identified data, they should have appropriate governance arrangements to manage privacy risks.

Agencies need to know what data they publish and who is responsible for it. They must also have confidence that their frameworks, policies and procedures manage re-identification risk adequately.

In this chapter, we consider how agencies use governance arrangements to manage re-identification risk in public datasets. For governance arrangements to be adequate, we expect agencies:

- capture all published de-identified datasets in a central register
- assign custodians or data owners to each de-identified dataset
- clearly define roles and responsibilities for releasing de-identified data
- outline a structured end-to-end process for the release of de-identified data
- provide sufficient guidance to decision-makers when releasing de-identified data
- regularly monitor and review de-identified data for changes in re-identification risk.

When looking at governance, we focused on arrangements relevant to managing re-identification risk. There are other risks agencies should manage when publishing data, such as commercial risks, however these fall outside the scope of this audit.

Conclusion

Both agencies have detailed governance arrangements for public data in general. However, only one agency has adequate guidance to assist decision-makers when releasing de-identified data. The other agency does not have sufficient guidance to support effective re-identification risk management. This means its governance arrangements do not fully manage the privacy risks of de-identified data.

Neither agency has appropriate governance arrangements to regularly monitor and review re-identification risk in de-identified datasets. Without these arrangements,

neither agency can be confident that risk management strategies remain effective over time.

Registers and custodians

To manage the privacy risks of public data, agencies must know what de-identified data they publish, and who is responsible for it.

Agencies should maintain accurate records of published de-identified datasets, ideally in a data register. These records should also assign a data custodian to each dataset. Data custodians are responsible for publishing and maintaining datasets over their life cycle.

Both agencies maintain current registers that capture public datasets and their custodians, as outlined in Figure 6 below.

Figure 6

Register features

	Agency 1	Agency 2
Captures released de-identified datasets	✓	✓
Assigns custodians to all datasets	✓	✓
Highlights de-identified data	X	X

Neither agency maintains a dedicated register of de-identified datasets nor has an easy way to locate de-identified data in their general register (for example, a de-identified data flag). Given the inherent risk of de-identified public data, maintaining a dedicated register of, or a way to easily locate, de-identified data would assist agencies to target their risk management strategies.

One agency has a systematic method for reviewing and updating its register to maintain accuracy. The other has an informal process. A systematic review method is better practice, as it provides greater assurance that registers are current and accurate.

Recommendation 2

We recommend all government agencies that publish de-identified data:

Assign a data custodian to each published de-identified dataset and capture this information in a register.

Roles and responsibilities

Simply assigning custodians to datasets does not manage privacy risks. Each actor in the de-identification and publication process must understand their roles and responsibilities. Policies and procedures should outline who is responsible for approving, releasing and maintaining de-identified datasets. This information should be easy to locate and clearly articulated.

Both agencies clearly define roles and responsibilities for public data releases, as outlined in Figure 7.

Figure 7

Key roles and responsibilities

Agency 1



Data custodian

- Identify datasets and determine suitability for release
- De-identify datasets where required
- Maintain and update data assets



Authorising officer

- Approve datasets for release
- Obtain endorsement from divisional head



Information and engagement team

- Implement open data strategy
- Provide guidance on open data requirements

Agency 2



Source: Office of the Information Commissioner

Both agencies clearly document who is authorised to release de-identified data. In both cases, the final approval to release de-identified data sits with a sufficiently senior decision-maker. An information management committee has oversight of public data in both agencies, which is a feature of effective program governance.

Governing policies

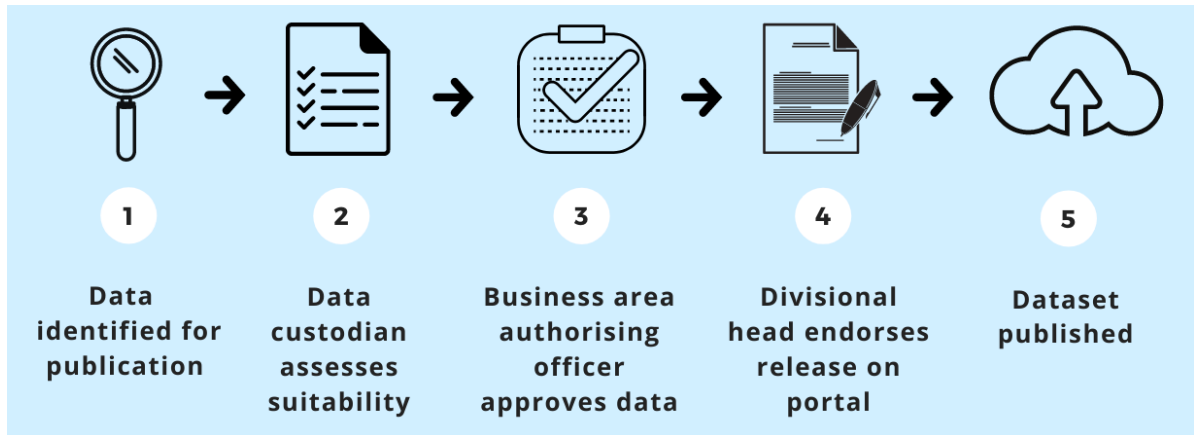
To manage the risk of de-identified data, agencies should document the process for releasing data on public platforms. Policies and procedures should be clear, accessible and sufficiently detailed.

Both agencies have high-level policies and procedures that govern public data releases. They also have an open data strategy setting the principles and objectives for their public data activities. Both agencies have an open data procedure document that outlines governance arrangements in sufficient detail. Figure 8 outlines how the agencies select, approve and publish public data.

Figure 8

Public data process

Agency 1



Agency 2



Source: Office of the Information Commissioner

One agency explicitly discusses privacy risks in its governing policies, stating the agency will not publish data that infringes on privacy. The other does not reference privacy in its governing policies.

Maintaining policies and procedures also supports effective public data governance. One agency's strategy and procedure documents are out of date. The agency advised it is updating these documents.

Guidance for decision-makers

De-identification is technically complex and can be high risk. Decision-makers need appropriate information to facilitate de-identified data releases. This includes guidance on de-identification techniques, re-identification risk and appropriate risk treatments. This guidance should be detailed, methodical and comprehensive.

Only one agency has sufficient guidance to support decision-makers, as outlined in Figure 9, noting both agencies can improve their guidance.

Figure 9

Decision-maker guidance features

	Agency 1	Agency 2
Prohibits the release of explicit personal information	✓	✓
Mentions de-identification as a strategy	✓	✓
Provides guidance on de-identification methods	✓	X
Warns that de-identified datasets can be re-identified	✓	X
Prompts users to think about risks in the wider data environment	✓	X
Requires users to consult with privacy or data experts	X	X
Requires users to document re-identification risk and de-identification techniques	X	X

While both agencies have guidance to assist decision-makers when releasing de-identified data, the level of detail varies considerably.

Agency 1

The agency has a checklist for decision-makers considering a public data release. This comprehensive document considers a range of relevant public data factors, including privacy. Decision-makers have access to the checklist and any comments from the data custodian and consulted parties when approving the release.

The checklist explicitly warns against releasing data containing 'sensitive information' and directs users to our guidance on privacy and de-identified data.¹⁶ While the checklist does not contain guidance on de-identification itself, the link to our material provides sufficient information on de-identification methods and re-identification risk.

The checklist could be improved. For example, it lacks a field for comments on privacy or de-identification considerations. These comments are necessary to fully inform decision-makers on the de-identification methods applied to the data and the relevant re-identification risk. The checklist also contains unclear terms, such as 'personal data' and 'sensitive information' rather than 'personal information' as defined in the *Information Privacy Act 2009*.

Agency 2

The agency has an assessment tool used to shortlist data that may be suitable for publication. It considers the costs involved with preparing the data for publication, and the public utility of the data. The tool prompts users to consider privacy and refers to the *Information Privacy Act 2009*.

The assessment tool only notes that personal information can be removed or de-identified to allow for publication. It is silent on de-identification methods and re-identification risk and does not link to any additional guidance to assist decision-makers.

The agency uses the assessment tool to produce an approval document for decision-makers. Relevant senior stakeholders endorse the document and approve release. While there is a work instruction about preparing approval documents, there is no requirement to discuss privacy risks or de-identification in the document.

While the agency has a structured decision-making process for public data in general, it lacks sufficient guidance on de-identification techniques and re-identification risk. For this reason, the guidance does not adequately support decision-makers when releasing de-identified data.

¹⁶ OIC guidance on Privacy and De-identified Data: <https://www.oic.qld.gov.au/guidelines/for-government/guidelines-privacy-principles/anonymity/privacy-and-de-identification>

Recommendation 3

We recommend all government agencies that publish de-identified data:

Implement and maintain policies or procedures that govern public de-identified data releases, including guidance to decision-makers.

Monitor and review procedures

The external data environment is rapidly changing, driven by advances in technology and an increase in public information. We expect agencies regularly review the re-identification risk of their published datasets. They should also monitor the data environment and conduct ad hoc risk reviews as necessary.

Policies and procedures should clearly outline who is responsible for reviewing re-identification risk, when these reviews should occur, and what they should consider. These procedures should follow similar principles to the decision-maker guidance.

To be effective, a re-identification risk review:

- revisits the original assessment and considers if risks have changed
- assesses the effectiveness of existing risk treatments
- develops and implements new treatments if necessary.

Sometimes, changing or emerging risks may require datasets be further de-identified or removed from public platforms altogether.

Both agencies have procedures to review and update public data. This includes reviewing general data quality and making periodic content updates. However, as outlined in Figure 10, neither agency has policies or procedures to monitor and review de-identified datasets for changes in re-identification risk.

Figure 10

Re-identification risk monitor and review procedures

	Agency 1	Agency 2
Monitors the external data environment	X	X
Regularly reviews datasets for changes in re-identification risk	X	X

Without these governance arrangements in place, agencies cannot have confidence that their risk management strategies remain effective over time.

Recommendation 4

We recommend all government agencies that publish de-identified data:

Monitor the external data environment and the effectiveness of risk treatments, and regularly review existing de-identified datasets for changes in re-identification risk.

4. Managing re-identification risk

Introduction

When publishing de-identified data, agencies must have a clear picture of the re-identification risk in their datasets and the wider data environment.

Agencies should understand the unique values in their data and how these could be linked to auxiliary information. From this, agencies can develop and apply appropriate de-identification techniques to decrease re-identification risk to an acceptable level.

In this chapter, we examine how selected agencies manage re-identification risk when publishing de-identified data. To effectively manage re-identification risk, we expect agencies:

- identify and assess re-identification risk before publishing de-identified data
- develop adequate strategies to manage re-identification risk, including appropriate de-identification techniques
- implement risk treatments before publishing de-identified data
- regularly monitor and review de-identified data for changes in re-identification risk.

We examined eight de-identified datasets across the two audited agencies.¹⁷ The de-identification techniques applied to these datasets and their apparent re-identification risk vary. Our findings are specific to the eight datasets discussed.

Re-identification risk analysis is technically complex. We engaged CSIRO's Data61, the data science and digital specialist arm of the Commonwealth Scientific and Industrial Research Organisation, to assist with this section of the report.

Data61 are experts in de-identification and re-identification risk analysis. They have specialised analytic tools that quantify the re-identification risk 'score' of de-identified data. We used these risk scores, and Data61's supporting analysis, to inform our findings in this chapter.

¹⁷ Where agencies have published multiple versions of the same dataset, or multiple datasets with similar attributes but different populations, we have treated these as a single dataset in this report.

Conclusion

While the agencies applied de-identification techniques to all eight datasets, neither could consistently demonstrate how it selected these techniques and determined they appropriately managed re-identification risk.

When assessing re-identification risk in the published data, we assigned relatively low risk scores to the datasets of one agency. The de-identification techniques used have effectively reduced the risk of re-identification to generally low levels.

The other agency has significantly higher risk scores, noting three datasets scored medium or above in our assessment. There is a significant risk of re-identification in these three datasets.

Neither agency monitors the data environment nor reviews the re-identification risk of the examined datasets. This means neither agency has assurance that its risk management strategies remain effective over time.

Assessing re-identification risk

Even when data appears de-identified, there is always a risk of re-identification. For this reason, agencies should conduct a detailed re-identification risk assessment for every de-identified dataset before publishing. They should document and keep this assessment to facilitate periodic risk reviews.

We detail the principles for a re-identification risk assessment in Chapter 2. Neither agency maintains adequate records of following these principles for the selected datasets, as outlined in Figure 11.

Figure 11

Re-identification risk adequately documented			
Agency 1		Agency 2	
Dataset 1	X	Dataset 1	✓
Dataset 2	X	Dataset 2	✓
Dataset 3	X	Dataset 3	X
Dataset 4	X	Dataset 4	X

Agency 2 has a detailed approval document for two datasets. This document:

- notes the raw data contains personal information
- specifies the attributes that are at risk of re-identification
- warns data may be linked with auxiliary information
- suggests appropriate de-identification techniques to treat risk.

This is a good example of managing re-identification risk when deciding to release data. By taking this approach, the agency has identified risks, considered the external data environment and recommended treatment strategies.

Recording this process and the outcome assures decision-makers that re-identification risk is managed. It also helps the agency when reviewing the risk and associated treatments over the data's lifecycle.

The other two datasets from Agency 2 have insufficient records of re-identification risk management:

- For one dataset, the approval document states '*no personal information about customers will be released*'. This indicates the agency considered some privacy implications. However, without more detail, it is unclear if Agency 2 has adequately identified and treated the risk.
- There is no approval document discussing re-identification risk for the other dataset.

While Agency 1 does not have records of re-identification risk analysis for the selected datasets, there is evidence of a structured approval process in place at the time it first published the data. The agency has limited records about de-identifying two of the selected datasets. However, these do not adequately demonstrate how re-identification risk was assessed or treated.

Recommendation 5

We recommend all government agencies that publish de-identified data:

Manage privacy when publishing de-identified data by adequately capturing, assessing and treating re-identification risk.

Analysing re-identification risk

As the agencies do not have consistent records explaining the de-identification techniques applied to the selected datasets, we analysed the data to determine how effectively they treated re-identification risk in practice. We engaged CSIRO's Data61 to assist with technical aspects of the analysis.

In this section, we use the term 'individual' to refer to people whose personal information is in the selected datasets. An 'antagonist' is a person who might seek to re-identify an individual. They could be a malicious actor or simply a 'nosy neighbour'.

The analysis assigns a weighted risk score to each dataset. It considers the number of unique entries that can be re-identified in each dataset and balances this against:

- the likelihood of an antagonist knowing there is information about an individual in the examined data
- the likelihood of an antagonist having the auxiliary information needed to re-identify an individual in the examined data
- the technical complexity of re-identification
- the new information an antagonist would learn in a re-identification event, and the sensitivity of that information.

The resulting risk score is on a scale of 'very low' to 'very high'. We consider a score of medium or above to represent a significant privacy risk.

To protect individuals whose information is in the selected datasets, we do not name the audited agencies or the datasets we examined. Instead, we describe the datasets only in general terms. We have deliberately changed or obfuscated certain data characteristics to further mask the identity of the datasets. Where we discuss datasets in detail, examples of re-identification events are hypothetical.

Agency 1

Figure 12 summarises the re-identification risk analysis for Agency 1.

Figure 12

Re-identification risk analysis: Agency 1

Dataset 1

De-identification techniques


- Removal and suppression

Risk factors

- Dataset contains vulnerable population
- 38 per cent of entries are unique with knowledge of two attributes

Mitigating factors

- Some suppression of sensitive attributes



A semi-circular gauge with five segments: green (very low), light green, yellow, orange, and red (very high). The needle points to the boundary between the yellow and orange segments.

very low very high

Risk score
medium to high

Dataset 2

De-identification techniques


- Removal and suppression

Risk factors

- Dataset contains vulnerable population
- 22 per cent of entries are unique with knowledge of two attributes

Mitigating factors

- Some suppression of sensitive attributes



A semi-circular gauge with five segments: green (very low), light green, yellow, orange, and red (very high). The needle points to the yellow segment.

very low very high

Risk score
medium

Dataset 3

De-identification techniques


- Removal and suppression

Risk factors

- Dataset contains vulnerable population
- 84 per cent of entries are unique with knowledge of two attributes

Mitigating factors

- Some suppression of sensitive attributes



A semi-circular gauge with five segments: green (very low), light green, yellow, orange, and red (very high). The needle points to the boundary between the yellow and orange segments.

very low very high

Risk score
medium to high

Dataset 4

De-identification techniques

- Removal

Risk factors

- Unnecessary unique identifiers
- 22 per cent of entries are unique with knowledge of two attributes

Mitigating factors

- Difficult to obtain auxiliary information
- Lack of sensitive data



Source: Office of the Information Commissioner

The first three datasets from Agency 1 have a significant risk of re-identification, with two datasets rated as medium to high risk. While we observed evidence of de-identification in these datasets, this was not always performed consistently or effectively.

Case study: Agency 1, dataset 3

This large dataset contains de-identified information about vulnerable individuals that access a particular government service.

We observed evidence of de-identification techniques in this dataset, such as removing personal information and suppressing sensitive attributes. However, the agency has not applied the suppression techniques effectively across all at-risk attributes in the data.

Are there unique entries in the data?

There are only a small number of attributes with unique values. However, when combining two attributes, a significant number of entries are unique. These attributes are approximate information about the individual's address, and the precise date they accessed the government service.

On a combination of these attributes, an overwhelming 84 per cent of entries in this dataset are unique. While the data contains specific dates, we also assessed uniqueness if an antagonist knew the month of access only. Over 27 per cent of individuals in the data remain unique with a combination of approximate address and month of access.

How likely is an antagonist to know an individual is in the data?

As this dataset is about all individuals who accessed a particular government service, it is possible an antagonist may know this as auxiliary information. For example, an antagonist may overhear the individual discussing the government service. This would confirm the individual's information is included in the data.

How likely is an antagonist to know necessary auxiliary information?

It is reasonably easy to obtain general information about an individual's address from auxiliary information, for example, social media. An antagonist who knows that an individual is in the dataset would likely also know, or could easily obtain, general information about where that individual lives.

To re-identify an individual, an antagonist would also need to know information about when the individual first accessed a specific government service. This would be difficult to obtain as auxiliary information unless the antagonist witnessed the individual accessing the service, or it was somehow disclosed by the individual.

What could an antagonist learn, how sensitive is it?

An antagonist could learn a significant amount of information about an individual, much of which is sensitive. This risk is amplified by the vulnerability of individuals in this data.

What is the overall risk of re-identification?

When considering the volume of unique entries, the sensitivity of the data, and the likelihood of an antagonist obtaining necessary auxiliary information, we assess this dataset to have a medium to high risk of re-identification.

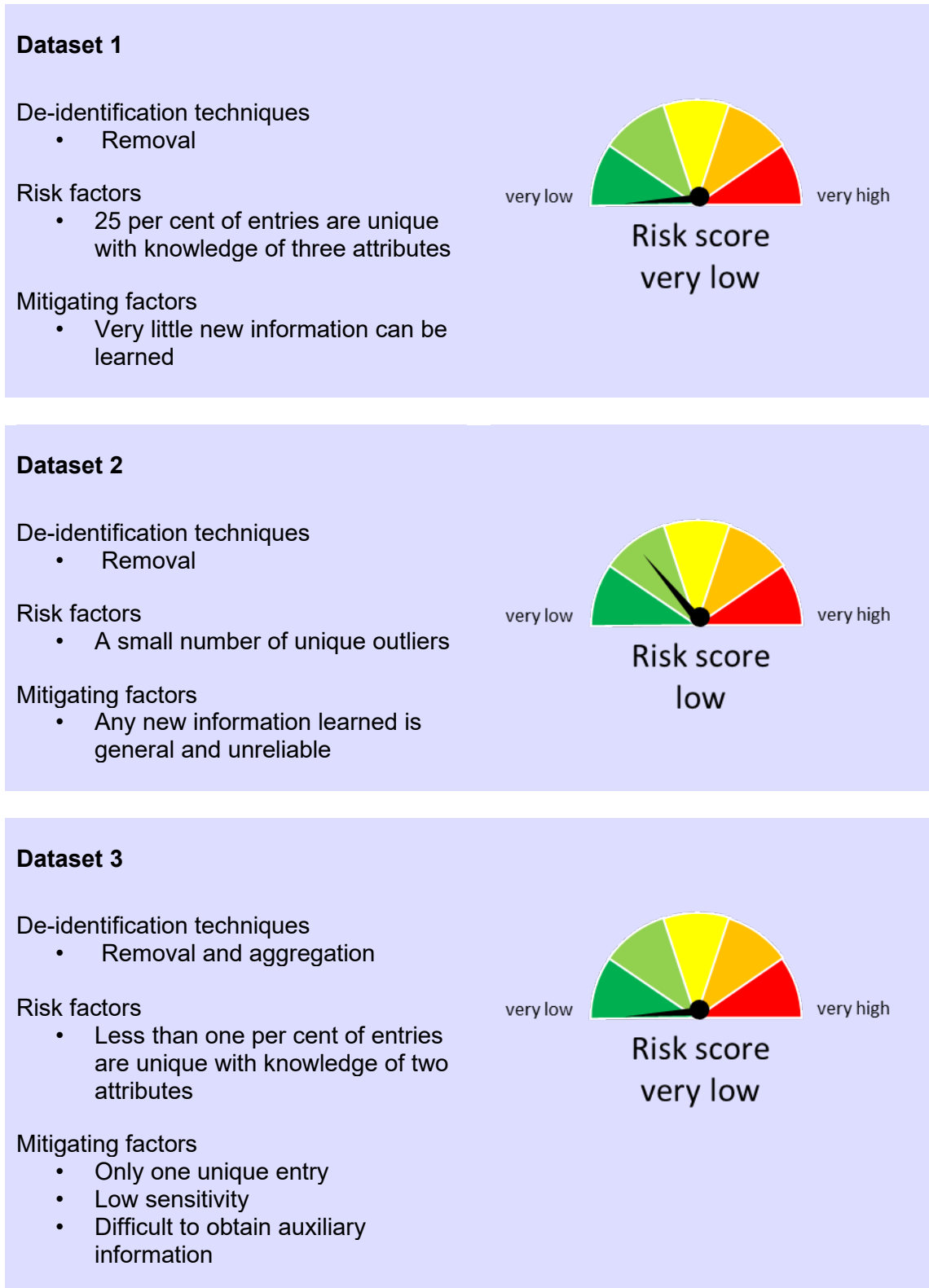
While difficult, it is entirely possible a motivated antagonist could obtain necessary auxiliary information and learn sensitive information about an individual in this dataset.

Agency 2

Figure 13 summarises the re-identification risk analysis for Agency 2.

Figure 13

Re-identification risk analysis: Agency 2



Dataset 4

De-identification techniques

- Removal, sampling and batching

Risk factors

- 18 per cent of entries are unique with knowledge of one attribute

Mitigating factors

- Difficult to obtain auxiliary information
- Hard to know an individual is in the data with certainty



Source: Office of the Information Commissioner

The four datasets from Agency 2 have reasonably low re-identification risk scores, with only one dataset rated as low to medium risk. This dataset contains one attribute with a high volume of unique values, although the risk is somewhat mitigated by the difficulty of knowing this attribute as auxiliary information.

No dataset contains inherently sensitive information. While the agency cannot provide records of how it decided which de-identification techniques to apply for all four datasets, the data is at relatively low risk of re-identification as published.

Case study: Agency 2, dataset 1

This large dataset contains basic de-identified information about reports made to the agency on a particular topic. Unlike the previous case study, this dataset does not relate to an inherently vulnerable population. It contains no sensitive information.

Are there unique entries in the data?

There are a small number of entries that are unique on one attribute or a combination of two attributes. When combining three attributes, a quarter of entries become unique. These three attributes are about the time, location and classification of the reportable event.

How likely is an antagonist to know an individual is in the data?

As the dataset is about all reports on a particular topic, the most probable antagonist is a person or organisation that has been involved in a reportable event. In this scenario, it is highly likely an antagonist knows an individual is in the data.

How likely is an antagonist to know necessary auxiliary information?

If the antagonist has been involved in the reportable event, they are likely to know the auxiliary information necessary for re-identification. The time, location and classification of the event would likely be known by any party associated with the event.

What could an antagonist learn, how sensitive is it?

While a quarter of the dataset can be re-identified if an antagonist knows the time, location and classification of the event, this exhausts all useful information in the dataset. There are no other attributes in this dataset that reveal meaningful information about the individuals involved in the event. It is unlikely an antagonist would learn anything new, or sensitive, from this dataset.

What is the overall risk of re-identification?

When considering the volume of unique entries, the sensitivity of the data, and the likelihood of an antagonist obtaining new or sensitive information, we assess this dataset to have a very low risk of re-identification.

While there may be some unique scenarios where an antagonist could learn new information, for example the time or location of an event, this risk is limited to a small number of outliers only. For the vast majority of this dataset, a meaningful re-identification event is unlikely.

Reviewing re-identification risk

Like any other risk, agencies need to monitor and review the re-identification risk of published de-identified data. Agencies should schedule and carry out re-identification risk reviews at regular intervals. It may also be necessary to review risk in response to external events, such as other data releases.

While both agencies have processes to regularly review public datasets for data quality, neither routinely monitors the data environment nor reviews de-identified datasets for changes in re-identification risk. Without regular reviews and ongoing monitoring, agencies cannot have confidence they are effectively managing re-identification risk over time.