Australian Government

Data Integration Involving Commonwealth Data for Statistical and Research Purposes: Risk Assessment Guidelines

(December 2013)

Contents

1 Introduction

1.1 Summary

Data integration increases the value of Commonwealth datasets by combining them to create more comprehensive information about Australia. Currently, data integration and the risks of integration are managed inconsistently across the Commonwealth.

The arrangements for the integration of Commonwealth data were proposed in 2010 to manage the risks of data integration. Their purpose is to encourage Commonwealth agencies to share their data for linking purposes in an effective and safe way. Consistent and robust processes were proposed to increase Commonwealth agencies' confidence in data integration projects, in particular the management of systemic risk.

The purpose of the risk assessment is to help Commonwealth agencies assess the level of risk of data integration projects as part of determining if a project should proceed and whether an accredited integrating authority is required to manage the integration project

The risk assessment is only one element that needs to be considered in making a decision on whether to proceed with a project. Data custodians need first to consider whether a project is appropriate and supported by the agency, taking the level of community acceptance, cost/benefit and political considerations of the project into account.

The risk assessment process involves the development of two risk assessments. The initial risk assessment assesses the risk of the data integration project against criteria specified by the Cross Portfolio Data Integration Oversight Board (the pre-mitigation risk assessment). A subsequent risk assessment assesses the residual risk after accounting for risk mitigation strategies (the post-mitigation risk assessment). The risk assessments are compiled either by a lead data custodian appointed by the data custodians for the project, or jointly by all the data custodians (as appropriate given the legislative environment), and may include input provided by one or more integrating authorities and data users. These two assessments are then submitted as part of project registration. The Oversight Board is able to review the risk assessments to ensure they are undertaken accurately and consistently across agencies, and provide advice to agencies on how they could manage risk.

These guidelines ensure that only truly 'high' risk projects require the use of accredited Integrating authorities (IAs), reinforcing the aim of the arrangements to encourage and enable greater sharing of Commonwealth data in an effective and systemically safe way.

1.2 Purpose

The purpose of the Guidelines is to help Commonwealth agencies assess the level of risk of data integration projects. This assists data custodians in determining if a project should proceed and whether an accredited integrating authority is required to manage the project. An accredited integrating authority is required if, even once risk mitigation strategies are put in place, there is a 'high' risk of harm to data providers (including persons, families, households or organisations who have contributed data) or a loss of public trust in the Australian Government or its institutions. [1]

1.3 Background

Data integration combines information from different data sources to produce new datasets. In 2010, the Secretaries Board endorsed a set of principles that govern the integration of Commonwealth data for statistical and research purposes, as well as a set of governance and institutional arrangements to support these principles.[2] This paper introduces the revised risk framework guidelines (the Guidelines) that enable Commonwealth data to be safely integrated.

1.4 Whether a project should proceed

The risk assessment is only one element that needs to be considered by agencies in deciding if a project should proceed. Other critical decisions are whether benefits outweigh costs, the level of community acceptance of a project and contextual issues. Contextual issues relate to the political, social and economic landscapes. A project also requires the support of data custodians. The appropriateness of a project may be influenced by whether the project is legally required, is a decision of government or meets a broad community need.

Privacy concerns have a strong bearing on community acceptance of a project. These concerns are influenced by the perceived sensitivity of the data being used. Public trust in government may be negatively impacted by projects integrating data people perceive to be sensitive or for a purpose they do not support. An assessment on the likely public perception of integration is therefore important. Transparent processes and community engagement will reduce the concerns the public have in integration projects. [3]

While a decision on whether to proceed with an integration project should be based on all of these elements, this risk framework focuses on assessing the risk of a breach of confidentiality and privacy. Breaches have negative consequences for public trust in government and for individuals and organisations that are affected by the breaches.

[1] Information about accredited integrating authorities can be found at:
http://www.nss.gov.au/nss/home.nsf/NSS/0E887A88A9224F8BCA2577F20016FE5D?opendocument

[2] National Statistical Service (NSS), Data Integration Landing Page:
http://www.nss.gov.au/nss/home.nsf/pages/Data+Integration+Landing+Page?OpenDocument

[3] Principles 4 & 7, High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes

2 Key concepts

Key concepts in the Guidelines fall under two key headings: the players in the process, and risk concepts in the Guidelines. The players are individuals and organisations that take part in the integration process, while the risk concepts are key terms used throughout the Guidelines.

2.1 Players in the process

2.1.1 Data custodian

- A data custodian:
    - is an agency accountable for managing the use, disclosure and protection of its data.
    - operates within legislative authority (where it exists) to provide data to integrating authorities for integration.
    - remains accountable for the data it has responsibility for throughout the integration process.[4]
    - approves data integration projects and appoints an integrating authority.
    - consults with other data custodians and the integrating authority to ensure all appropriate risk assessments are undertaken.
- Where there is more than one data custodian, a lead data custodian may be appointed by the data custodians to compile the risk assessments.
- In this framework, the term data custodian refers to either the sole data custodian, the lead data custodian where appointed, or custodians jointly where there are multiple custodians and no lead custodian.

2.1.2 *Integrating authority*

- An integrating authority:
    - is an organisation appointed by the data custodian to integrate two or more datasets, at least one of which is a Commonwealth dataset.
    - is the single agency[5] accountable for the sound conduct of the integration project, including ongoing risk management.
    - may also be a data custodian or data user.
    - may be an accredited Integrating Authority.[6]
    - may suggest changes to the risk assessments completed by the data custodian/data custodians.

2.1.3 Data user

- A data user:
    - is a person or an organisation that undertakes analysis of an integrated dataset.
    - may also be a data custodian and/or an integrating authority.

[4] The accountability of the data custodian is described in Principle 2 of the High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes.

[5] More than one integrating authority may be consulted or asked for a quotation on each project before it commences. However, only one integrating authority can have full responsibility for the way in which each data integration project is conducted, and this must be determined before the project commences.

[6] Further information on the accreditation process can be found at: www.nss.gov.au/nss/home.nsf/pages/Data%20Integration:%20Accredited%20Integrating%20Authorities

  - does not need to be a Commonwealth agency.
- There may be many data users involved in a project.

2.1.4 The Cross Portfolio Data Integration Oversight Board (The Oversight Board)

- The Oversight Board:
  - is responsible for the arrangements for the integration of Commonwealth data for statistical and research purposes on behalf of Secretaries Board.
  - provides strategic and collaborative leadership, supports effective governance and may provide advice to help manage the risks of particular data integration projects.
  - helps manage the systemic risk associated with conducting multiple data integration projects involving Commonwealth data through assessment of proposed risk mitigation strategies and the provision of advice.
  - endorses any changes or additions to the overall environment, including amendments to the principles or guidelines, or the development of new general tools to support integration or safe access to integrated data for statistical and research purposes.
- In practice, the Oversight Board
  - has ten working days following registration of the project and receipt of the risk assessment to raise any concerns about the project with the data custodians or integrating authority. These concerns relate to the management of systemic risks of data integration.
  - has no authority to approve or delay data integration projects. Approval is given by data custodians.
  - ensures that the risk mitigation strategies proposed when a project is registered are implemented. To ensure this, the Oversight Board may request a review of one or more of a data custodian's integration projects.
  - may work with data custodians and integrating authorities to improve their risk assessment processes.
  - can delegate its review functions.

- The Oversight Board will work with data custodians and integrating authorities to resolve any issues relating to unacceptably high systemic risks or inappropriately managed projects. If the issues cannot be resolved or managed to the satisfaction of the Oversight Board, then the Chair of the Oversight Board will engage in direct discussion and negotiation with the agency head of each data custodian that is party to the project to resolve the matter. Where there is a conflict of interest for the Chair of the Oversight Board to engage in direct discussion with the head of each the agency concerned, the matter will be referred to another member of the Oversight Board and where resolution cannot be achieved, to the Secretaries Board.

2.1.5 Data provider

- is an individual, household, business or other organisation which supplies data to a data custodian.

2.2 Risk concepts

2.2.1 Risk Assessment Guidelines (the Guidelines)

- The Guidelines provide a platform to assess the risk of harm to a data provider and the risk of a reduction of public trust in the Australian Government and its institutions as a result of a breach.
- Data custodians can decide that the assessment guidelines on risk dimensions are not valid for their particular context. However, deviations from the assessment guidelines must be explained in the risk assessment.

2.2.2 Breach

- A breach is "when personal information or the confidential information of an organisation held by an agency or another organisation is lost or subjected to unauthorised access, use, modification, disclosure, or other misuse."[7]
- For example, a breach occurs when:
    - a USB with unconfidentialised data is left on a train and viewed by an individual not authorised to see the data.
    - data is poorly confidentialised and exposes information about a data provider.
    - an aggregate table is published with small cells that allow identification of data providers. For example consider a table that cross-tabulates the number of recipients of carer payments by sex, suburb and income, which contains a cell showing only one male recipient living in Glebe. Any users who know a male in Glebe who receives a carer payment will be able to use the table to determine his income.
- If there is legislation which impacts on the dissemination and management of data, the more stringent understanding should apply.

2.2.3 Risk Assessment

- The risk assessments are undertaken by the data custodian and involve several steps:
    1. a pre-mitigation risk assessment is conducted following consultation with other stakeholders.
    2. mitigation strategies are developed. This step may involve consultation with the integrating authority and data users.
    3. a post-mitigation risk rating is calculated which determines whether an accredited Integrating Authority is required.
- If the post-mitigation risk rating for a project is 'low' or 'medium', the use of an accredited Integrating Authority is optional.

[7] One definition of a breach is given at http://www.oaic.gov.au/privacy/privacy-resources/privacy-guides/data-breach-notification-a-guide-to-handling-personal-information-security-breaches#_Toc301281661. This has been expanded to include organisational information.

- The risk assessments and post mitigation risk rating for a data integration project are submitted to the Oversight Board via the project registration process to ensure that they have been completed appropriately. Registration occurs after a project has been approved by data custodians and agreements signed.

2.2.4 Likelihood of a breach

- Likelihood is the measurement of the potential for a breach to occur.[8]
- It assesses the parts of a project in which a breach is possible.
- For example, the security of an agency's IT systems will determine how likely a breach is to occur due to hacking.

2.2.5 Consequences of a breach

- Consequence is a measurement of the potential outcome of any breach.
- It includes harm to a data provider, including humiliation, or a reduction in public trust in the Commonwealth's ability to store and protect sensitive data.[9]

2.2.6 Risk review

- There are two types of review outlined in the Guidelines:
    - An initial review within the ten day period of the Oversight Board receiving a risk assessment for a particular project, and
    - A review of a data custodian's completed integration projects.
- The ten day review allows the Oversight Board to provide advice and guidance on specific projects. The advice may be in the form of concern around a particular project, or suggestions of ways to manage risk. This advice should focus on minimising the systemic risks of a project.
- The review of completed integration projects involves an assessment of public reaction to the project and how public understanding and acceptance was managed throughout the project, as well as an assessment of the application and effectiveness of the mitigation strategies.
    - Such a review will be conducted by a party:
        - approved by the data custodian;
        - that is at arm's length to the risk assessment process, but need not be external to the agency; and
        - that has the ability to provide an objective expert assessment of whether the mitigation strategies initially proposed were implemented.

2.2.7 Risk mitigation strategies

- Mitigation strategies attempt to minimise the risk of a project. In this risk framework, mitigation strategies will mostly act on the likelihood of a breach.
- Multiple mitigation strategies may be required to lower the risk of a breach.
- It is possible that mitigation strategies may lower one risk while increasing another.

[8,] These definitions draw on Standards Australia/Standards New Zealand's definitions of likelihood and consequence in the *Risk management – principles and guidelines* (2009).

[9] These concepts are further discussed in the NHMRC *National Statement on Ethical Conduct in Human Research*.

2.2.8 Public trust in the Government

- Public trust in the Government and its institutions is the degree to which individuals and organisations trust the Commonwealth, state, territory and local governments to manage their data.
- Many people do not distinguish between government agencies. Therefore, the actions of one agency affects people's trust in all government agencies.
- Public trust also impacts on the Australian Statistical System in general.
- Public trust affects how likely it is that individuals and organisations will participate in research conducted by government agencies.

2.2.9 Confidentialised data

- Confidentialised data is individual[10] or micro level unit information[11] that has been de-identified and has had other information removed or modified to reduce the risk of a data provider being identified.

2.2.10 Internal data integration

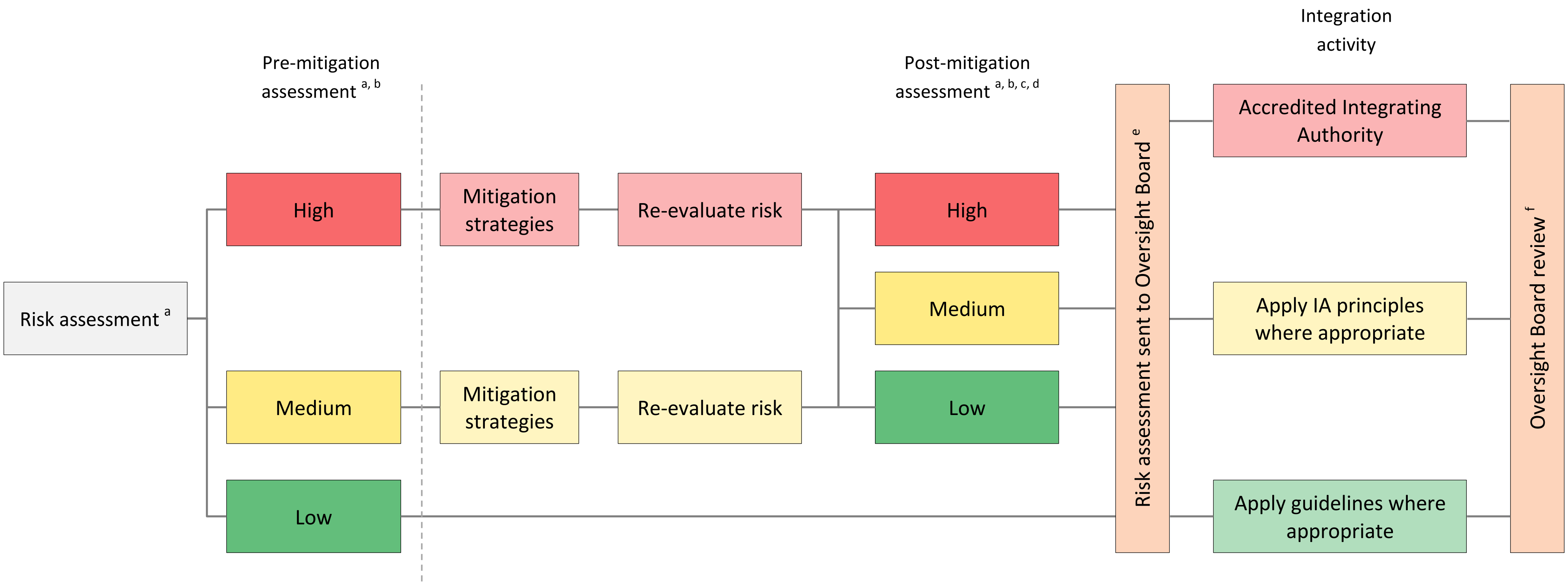
- Internal data integration is integration that is undertaken completely within an agency, not provided to external agencies or researchers and where the agency is also the custodian for all the Commonwealth datasets being integrated.
- Work being undertaken internally does not automatically lower the risk of a breach. However, there are many elements of internal work that are likely to be effective risk mitigation strategies. For example, the IT environment is likely to be more secure and legislative penalties for a breach high.
- The mitigation strategies should be included in the post-mitigation risk assessment.

[10] That is, a person or an organisation

[11] That is, a family, household, community, or organisation group.

3 Risk assessment process

Process and responsibilities


Integration activity
Pre-mitigation assessment [a, b]
Post-mitigation assessment [a, b, c, d]
Risk assessment [a]
High
Medium
Low
Mitigation strategies
Re-evaluate risk
Mitigation strategies
Re-evaluate risk
High
Medium
Low
Risk assessment sent to Oversight Board [e]
Accredited Integrating Authority
Apply IA principles where appropriate
Apply guidelines where appropriate
Oversight Board review [f]

a. Before a risk assessment is completed, the public's acceptance of the project should be considered and only projects with acceptable risk should proceed.

b. The data custodians are responsible for ensuring the completion of all risk assessment documentation.

c. Data custodians are responsible for approving a project once the risk assessment has been completed.

d. Integrating authorities and data users may provide expert advice to the data custodians on the likelihood of a breach, and the processes that exist in their organisations to protect data.

e. The Oversight Board has ten working days to raise any concerns about the project with the data custodians and/or integrating authority.

f. The Oversight Board may (at their discretion) request a review of completed projects to ensure that the risk mitigation strategies proposed were implemented. The data custodians appoint the reviewer.

3.1 Process

3.1.1 Pre-mitigation risk assessment

The first stage of the process involves the data custodian undertaking consultation with stakeholders. Key stakeholders are data custodians, integrating authority and data users. A data custodian compiles the pre-mitigation risk assessment. This assessment follows the assessment guidelines set out in section 4 of this paper. In compiling this assessment, the data custodian needs to consider whether the information in section 4 is relevant to the context of the particular project. For example, a project that is working with data that are culturally sensitive may need to consider the cultural impact of the research.

3.1.2 Mitigation strategies

Mitigation strategies lower the risk of a data breach. To analyse mitigation strategies, their impact on the overall risk of a project is assessed over the duration of the project. The data custodian leads this work, although other key stakeholders may play an active role in this process. The positive and negative effects of any mitigation strategies need to be assessed. For example, using expert contractors to undertake the integration decreases the technical complexity risk, yet it increases the managerial complexity of the project. As the aim of the Guidelines is to enable the research while managing risks, there should be a focus on what satisfies data custodians that data will be managed appropriately.

3.1.3 Post-mitigation risk assessment

Once it has been decided which mitigation strategies will be used, a post-mitigation assessment is compiled by the data custodian. This justifies the mitigation strategies and explains how they lower the risk of a breach. The risk assessments are then submitted to the Oversight Board as part of the project registration process. The post-mitigation risk assessment may need review as the project progresses. Risk assessment is an ongoing responsibility for the data custodians and integrating authority. If the project's risk changes significantly during the life of the project, then the risk assessment will need to be updated by the integrating authority in consultation with the data custodians and data users.

3.1.4 Integrating decision

The risk assessment process establishes whether an accredited integrating authority is required. If a project remains 'high' risk after mitigation strategies have been applied, then an accredited integrating authority is required. Where appropriate, the data custodian may assist the integrating authority in applying the best practice for integration.[12]

3.1.5 Next steps

The next steps for the project involve data custodians making a final decision to proceed with the project, based on public benefit and acceptance considerations, the risk assessment and the ability to mitigate that risk. If the data custodians approve the project, they will appoint an integrating authority, which may also be one of the data custodians, who will be responsible for the ongoing risk management of the project. The integrating authority, in consultation with the data custodians and data users, will finalise the details of the project

[12] The *Best Practice Guidelines* will have further information on best practices for integration.

and prepare agreements to formalise relationships between the parties involved in the integration project where required.

3.1.6 Oversight Board review

Registration of a data integration project occurs after the project is approved by the data custodians and agreements signed. The risk assessment for an integration project must be submitted to the Oversight Board when the project is registered.

The Oversight Board has ten working days to raise any concerns about or suggest improvements to the project with data custodians and the integrating authority. This step is not a road block to a project and the project may proceed immediately. The Oversight Board will work with data custodians and integrating authorities to resolve any issues relating to unacceptably high systemic risks or inadequate risk mitigation. The Oversight Board may delegate this role to another body. If the issues cannot be resolved or managed to the satisfaction of the Oversight Board, then the Chair of the Oversight Board will engage in direct discussion and negotiation with the agency head of each data custodian that is party to the project to resolve the matter. Where there is a conflict of interest for the Chair of the Oversight Board to engage in direct discussion with the head of each the agency concerned, the matter will be referred to another member of the Oversight Board and where resolution cannot be achieved, to the Secretaries Board.

The Oversight Board may also change the risk assessment process in future, in consultation with stakeholders, if it finds that the process does not accurately assess the true risk of projects.

4 Risk assessment guidelines

4.1 Introduction

The risk assessment guidelines provide assistance to organisations who are involved in an integration project in assessing the risk of a breach. This section outlines key components of the Guidelines, some thresholds of risk and some mitigation strategies. These are designed to be integrated into Commonwealth agency approval processes to assist with decisions about integration projects. A robust and consistent risk assessment process will ultimately increase confidence in data integration as a method of maximising the value of Commonwealth data. However, the Guidelines can be adapted to suit an agency's context, as risks will vary by situation and organisation. For example, the same project undertaken by Department of Social Services (DSS) may have a different level of risk than one done by Department of Health because of the pre-existing conditions within the organisations.

Further research needs to be done to provide better guidance on effective mitigation strategies and measuring these risks. As the Guidelines are designed to evolve, changes will be made as the process matures.

4.2 Dimensions of risk

The Oversight Board approved eight dimensions of risk in August 2011. The dimensions are:

- sensitivity,
- size,
- nature of data collection,
- technical complexity,
- managerial complexity,
- duration of project,
- how the data is to be linked, and
- nature of access.

Originally these dimensions were suggested to describe the nature of a project and have since been used to assess the risk of a project. As they describe the whole nature of a project, some are less relevant to the risk of harm to a data provider and of a loss of public trust in Government. The definitions that follow refine these dimensions of risk to account for this redundancy.

- Nature of data collection has been redefined as consent, as this is the most relevant component.
- The definition of technical complexity has been tightened to focus on the challenges of appropriately confidentialising information.
- The size dimension has been split to refer to the number of quasi-identifying variables, and the amount of other, less identifying information about a data provider in a dataset.
- The 'how the data is to be linked' dimension has been dropped as it does not significantly affect the likelihood of a breach.

The following sections offer guidance on how to assess risk by:

- providing additional guidelines to determine whether a project can be considered ‘high’, ‘medium’ or ‘low’ risk,
- exploring the importance of each of the dimensions of risk, and
- discussing some additional risks that relate to each dimension.

It is intended to provide a set of ‘rules of thumb’ rather than definitive advice. The data custodian may use their judgment if they justify the reasons for their departure from the following guidance.

An adverse public perception of a data integration project, regardless of whether or not there is likely to be a data breach, may lead to a considerable loss in public trust in all government data collection activities, having a broad impact on all government departments. This consequence of this systemic risk must be considered before it is decided to proceed further with a project.

4.2.1 Consequence of a breach

Sensitivity

Sensitivity assesses the effect of a breach on the data providers,.

The National Health and Medical Research Council’s *National Statement on Ethical Conduct in Human Research*[13] provides a good framework for assessing the sensitivity of the data in relation to persons. The main elements of a risk of harm are:

- physical harm,
- psychological harm,
- economic harm,
- social harm,
- legal harm, and
- devaluation of personal worth.

The National Statement focuses on individuals, but many of the concepts are relevant to organisations.

Most ethical risk frameworks assess whether the risk might affect populations over which the data user has additional duty of care obligations. These populations are usually children, people with mental illness, and people with cognitive or intellectual impairments. This may also extend to other small population groups where their information may be sensitive, such as Aboriginal or Torres Strait Islander populations or groups from particular ethnic or religious backgrounds. A way to account for this additional duty of care is to increase the risk rating of a project if these populations are likely to be affected. For example, an initial sensitivity rating of ‘low’ may be increased to ‘medium’ if a project included children. However, there are exceptions to this rule. For example, some information about school

[13] The NHMRC *National Statement on Ethical Conduct in Human Research* can be found here: http://www.nhmrc.gov.au/guidelines/publications/e72

children may be less sensitive if there is little variation in the dataset, or where the subject/topic is inherently of low sensitivity (for example, participation in sport).

The Guidelines take a different position on the classification of harm in comparison to the *National Statement on Ethical Conduct*. The Guidelines classify harm risk ratings as the following:

- 'High' consequence involves a foreseeable risk of serious harm to data providers in any of the main elements of harm.
- 'Medium' consequence involves a foreseeable risk of any harm to data providers.
- 'Low' consequence involves no foreseeable risk of harm.

The reason for this is the need to focus the risk management effort of the Oversight Board on those projects that pose the greatest risk to public trust and thus have the greatest potential to undermine the value of Commonwealth data as a strategic asset. Many projects carry some risk of harm in the event of a breach, but in general this risk must be managed by those directly involved (the data custodians, the integrating authority, and the data users).

Consent

Consent is the component of the nature of data collection that impacts on public trust in the Government. If an agency has informed consent from the provider, the consequences of a breach may be lower. Consent may lower the consequence as data providers are aware, or partially aware, of the research being undertaken and the risks of participating in the research. To be able to consent, there must be an option to opt-out for the data provider.[14]

- 'High' consequence involves no consent, or coerced consent.
- 'Medium' includes partially-informed consent.
- 'Low' consequence exists when informed consent has been obtained and the risks of integration have been explained and are understood.[15]

There is much literature on the notion of consent. While most focuses on the notion of an individual, the concepts that apply to an individual can easily be applied to organisations.

Amount of information about a data provider

The number and nature of variables containing information about a data provider in a dataset affects the consequence of breach. The National Statistical Service's (NSS) *Best Practice Guidelines for Integration* will provide a number of examples on how to mitigate this

[14] Opt-out refers to the concept of being able to decline to be involved in a research project without fear of repercussions. For example, those on government payments would have to be reassured that by not consenting for their information being used for research purposes that their current and future potential to claim payments will not be jeopardised. It is not enough to say that a payment is voluntary and therefore if they do not want their information used they can choose not to receive the payment.

[15] The Australian Communications and Media Authority's (ACMA) paper *Community research on informed consent: Qualitative research report* (2011) notes that "[customers] often gave 'consent' but claimed that in reality it was not always 'informed consent', as…they often provided consent without a full understanding and comprehension of the terms and conditions of the agreement."

risk. For example, the separation principle limits the number of variables in a given dataset by splitting datasets into linking variables and analysis variables.

Rating consequence risk

A project is assessed as having a ‘high’ consequence risk if:

- The level of consent or sensitivity of data risks have been rated as ‘high’, or
- The level of consent or sensitivity of data risks have been rated ‘medium’ and the risk due to the amount of information about a data provider rated ‘high’.

A project with ‘low’ consequence risk has no dimensions rated ‘high’ and a maximum of one dimension assessed as ‘medium’.

If the amount of personal information is rated as ‘high’, but the other two are ‘low’ then the overall rating is ‘medium’.

All other combinations of risk are rated as ‘medium’.

The sensitivity of data, and consent have been weighted equally as they are both can have very serious consequences. The amount of personal information is covered to a certain degree by the sensitivity of the data dimension.

4.2.2 Likelihood of a breach

Likelihood of identification

Not all variables are equally identifying in nature. Some variables are subjective (for example, have you been depressed most of the time for more than a month). Some have low visibility to others (for example, do you play golf). Other variables may be easy to assess objectively (for example, do you work as a teacher). Some are highly identifying (for example, name and address). The number and nature of the variables contained in a dataset affect the likelihood of identification. The inclusion of some variables, known as quasi-identifying variables, quickly increases the probability of spontaneous identification of records and identification through list matching[16]. For example date of birth and country of birth are quasi-identifying variables. In isolation they are not identifying; however, in combination they may be unique to an individual. In many circumstances, three quasi-identifying variables in combination may be enough to enable identification. For example, a male working in DSS that was born on 29 August 1987 and lives in Lyneham is likely to be identifying. Additionally matching these details against a publicly available dataset may provide ample information to identify persons or organisations within the de-identified dataset. However, removing one of these pieces of information may remove the identifying nature. The likelihood of a breach rises quickly with the number of quasi-identifying variables, especially where datasets are to be released at a unit record level, for example in a confidentialised unit record file (CURF). This risk is:

[16] List matching is where a user has access to another source of data, such as an external administrative dataset, and attempts to match the two datasets using common data items.

- ‘High’ where there is a high probability that an individual can be identified using a combination of quasi-identifying and other variables on a dataset.
- ‘Low’ where it is unlikely that quasi-identifying and other variables can be combined to identify an individual.

Technical complexity

The technical complexity of a project affects the likelihood of a breach. It includes the complexity of confidentialisation and of methodology.

- ‘High’ risk involves complex data that is likely to be published as a CURF, or multiple aggregate tables that require consequential confidentialisation.
- ‘Low’ risk involves publishing simple aggregated data with basic confidentialisation required.

Managerial complexity

Data governance becomes more complex as the number of people and organisations involved in an integration project increases. This potentially leads to diminishing control over data management practices and therefore increases the risk of a breach.

The more organisations involved, the harder it is for the data custodians to influence the practices of these organisations. The more people involved, the higher the risk of data being leaked or used inappropriately.

- ‘High’ managerial complexity risk would have:
  - four or more agencies involved in the process, or
  - thirty or more staff directly involved in the integration.
- ‘Low’ managerial complexity risk would have:
  - only one agency involved, and
  - fewer than ten staff directly involved in the integration.

Duration of project

The longer data are stored following a project, the more likely a breach becomes. Similarly, the longer a project runs, the more likely a breach becomes. There are two reasons for the increased likelihood of a breach, data storage and documentation.

Data access that is poorly controlled, or data that are exposed to external attack, are poorly stored. The longer the data exists in this storage, the more likely it is that one of these deficiencies will be exposed and a breach will occur.

Poor metadata, or data documentation, may lead to new staff publishing data without appropriate confidentialisation, or using data inappropriately.

- A project with ‘high’ duration of project risk would:
  - retain data for more than three years, or
  - run for more than three years.

- A project with ‘low’ duration of project risk would:
    - destroy data on the completion of the project, and
    - run for less than one year.

Nature of access

The nature of access is concerned with the quality, consistency and coverage of governance and controls placed around access to data at all stages of the integration project.

- Unrestricted or unaudited access is a ‘high’ risk.
- ‘Low’ risk requires:
    - access to be granted on a demonstrated ‘need to know’ basis, and
    - the separation principle to be applied, and
    - regularly audited and restricted access.

Rating likelihood risk

It is difficult to weight the likelihood dimensions of risk, as the importance and impact of these dimensions depends on the context in which they are applied. As a guide the overall likelihood risk is

- ‘high’ if three or more likelihood dimensions have been assessed as ‘high’, and
- ‘low’ if no dimensions are rated ‘high’ and less than three are rated medium.

Mitigation strategies, especially for work undertaken internally within a Commonwealth agency, should reduce the likelihood risk considerably.

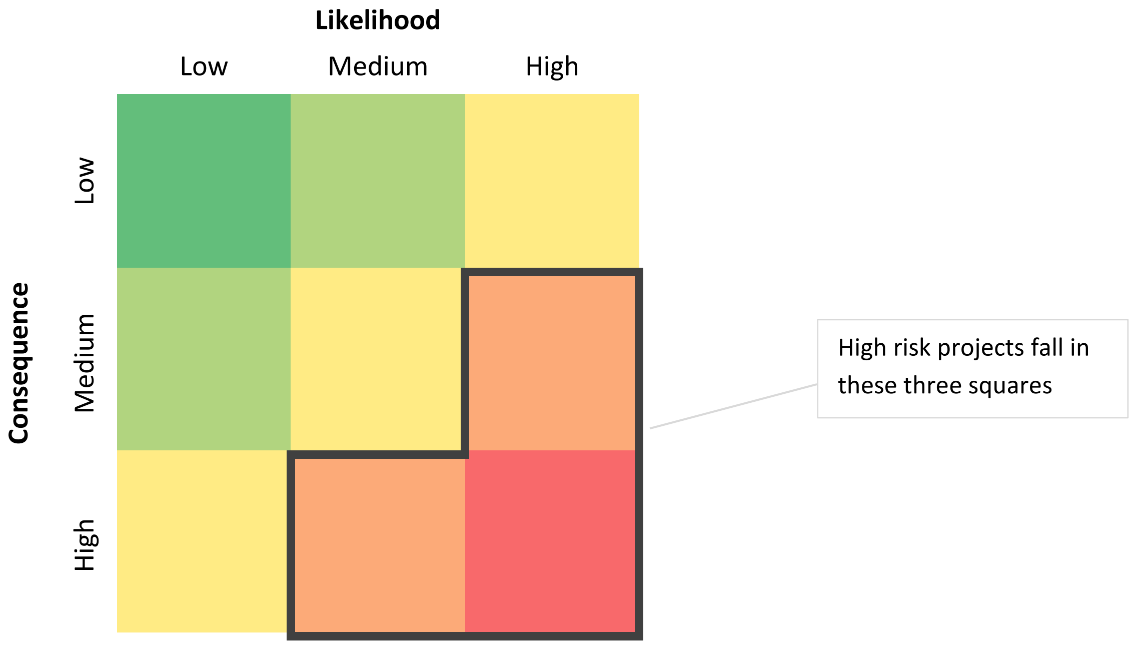
4.3 Relevant legislation

Legislation influences the way data are used and shared. Legislation that broadly covers data dissemination includes:

Privacy Amendment (Enhancing Privacy Protection) Act 2012
Privacy Act 1988
Privacy and Personal Information Protection Act 1998 (NSW)
Health Records and Information Privacy Act 2002 (NSW)
Information Act 2002 (NT)
Information Privacy Act 2009 (Qld)
Personal Information and Protection Act 2004 (Tas)
Information Privacy Act 2000 (Vic)
Freedom of Information Act 1992 (WA)
Industry Research and Development Act 1986
Pooled Development Funds Act 1992
Venture Capital Act 2002

However, there are many specific pieces of legislation that govern the use of data. An example of specific legislation is the *Social Security (Administration) Act 1999* which governs the way data collected by Centrelink is used and shared.

4.4 Likelihood and consequence risk matrix



High risk projects fall in these three squares

The likelihood and consequence risk matrix assists in determining the overall risk of a project. The overall risk rating is determined by the combination of the likelihood and consequence risk ratings. Once this rating is known, risk mitigation strategies can be identified and applied.

4.5 Mitigation strategies

There are many possible mitigation strategies that can be applied to reduce the likelihood risks. However, there are very few mitigation strategies that can be applied to the consequence risk dimensions without changing the scope of the project. For example, a dataset with highly sensitive data can have the sensitive data removed; however, this changes the project as the component data is now different.

4.5.1 Risk mitigation examples

Data labs for external data users

Requiring data users to use secure data labs ensures that the IT environment is more secure, limiting the potential for loss or theft of the data.

Legislative penalties

The data supplied by data providers may be subject to the Privacy Act and/or protected by one or more confidentiality/secrecy provisions that govern the management of that data. Data custodians, integrating authorities and data users are obliged to comply with the

Privacy Act and the confidentiality/secrecy provisions in relevant legislation governing the collection, use and disclosure of the information.

Elements of the integration Best Practice Guidelines

Some elements of the integration *Best Practice Guidelines* can be applied more easily than others, and they can be applied to different extents. For example, applying the separation principle may be costly and may be an operational challenge. However, separating a dataset into linking and analysis variables is relatively straight forward and reduces the size of the datasets. Therefore, if one of the datasets is compromised, only a subset of the information is made public, and harm may be averted. The consequence of the breach of one of these datasets is therefore lower than the breach of the combined dataset

Experienced data integrating staff

Using experienced staff ensures that processes are able to run more smoothly and efficiently. They will also be more likely to have an understanding of the data governance that applies to data integration and the purpose of the governance. Therefore, there is a lower risk of breaches resulting from negligence or ignorance.

This is by no means an exhaustive list of mitigation strategies. However, most mitigation strategies will impact on the IT environment, staff accountability and organisational procedures. As the risk assessment process matures, more mitigation strategies will become apparent. Choosing which to implement will in general be up to the data custodian and integrating authority involved in a data integration project. There are many mitigation strategies to consider for implementation. The key element of the post mitigation risk assessment is about how, and how much, the mitigation strategies proposed will lower the overall risk of a data integration project. The justification needs to satisfy the data custodians, integrating authority and, ultimately, the Oversight Board.

Appendix A – Single agency 'high' risk project case study

1. Purpose

The purpose of this appendix is to demonstrate through a case study, the process of determining the post mitigation risk rating of data integration projects using the draft risk framework guidelines. A practical example is provided in this document with both pre- and post- mitigation risks.

2. Definitions

For the purpose of classifying data integration projects, the following definitions are used:

Project Types	Definition
Single agency	There is only one Commonwealth data custodian involved in the data integration project.
Multiple agencies	There is more than one Commonwealth data custodian involved in the data integration project.
Non-Commonwealth	There is one or more non-Commonwealth data custodian involved in the data integration project.

3. Risks at different stages of the project

There are varying levels of risks associated with the different stages of a data integration project. Broadly, the stages include extraction, file transfer, linkage, analysis, publication, storage and destruction. Not all stages will be applicable to all data integration projects. For example, a data integration project may retain the linked dataset indefinitely. Therefore, the destruction stage may not be applicable.

Mitigation strategies can be applied at various stages to reduce the likelihood of a breach occurring. The data integration project can be designed in such a way that even if the consequence of a breach occurring is high, the likelihood is reduced such that the project does not require an accredited Integrating Authority. For example, even if the data required for the integration project are sensitive, using the separation principle may mean that the project does not necessarily require an accredited Integrating Authority.

4. Application of the framework through a selected case study

The following section provides a case study of pre- and post-mitigation assessments. The final risk assessment can assist in determining whether an accredited Integrating Authority is required or not.

The selected case study is:

#	Type of case study	Project Name
4.1	High Risk – Single agency	Client Data Collection (DSS)

Case studies from other agencies will be included later. This will help ensure the framework continues to be developed in a way that is fit for all situations.

4.1. High Risk – Single agency

Client Data Collection (CDC) data integration project (DSS)

Department of Social Services (DSS) proposed Client Data Collection (CDC) is an example of a high risk single agency project. DSS is the sole data custodian. In this case, DSS is also the integrating authority.

The purpose of the CDC project is to enable better monitoring and research within and between DSS programs and payments data. This project is still in the planning stage. The following strategies are hypothetical and subject to change.

4.1.1. Pre-mitigation Assessment

There are three dimensions that influence the consequence of a breach (from the eight risk dimensions agreed upon). The table below outlines the consequence (or impact) on individuals if there is a breach.

Dimension	Impact	Comments
Sensitivity	High	The project involves integrating all DSS programs and payments data. Information collected includes educational background, health status, income level and much more. The data is considered to be highly sensitive. If leaked, there is the potential to cause harm to individuals and the Commonwealth Government as a whole.
Consent	High	Some of the programs directly obtain consent[17] from clients. However, the majority of them do not. Generally, both programs and payments data are collected as an administrative by-product.
Amount of information about a data provider	High	There may be twenty or more variables with different personal information about a data provider.
Pre-mitigation Consequence Assessment	**High**	

There are five dimensions that influence the likelihood of a breach (from the agreed eight risk dimensions).

Dimension	Rating	Comments
Managerial complexity	Low	There will only be one agency involved in this project. However, a considerable number of internal stakeholders will be part of the project team. The number of DSS staff directly involved in the integration is fewer than ten.
Nature of access	Low	Restricted. Access granted to approved staff and access control to be reviewed regularly. The separation principle is applied.
Duration of the	High	Data is proposed to be retained for more than three years.

[17] Here clients gave consent for the information to be collected for statistical purposes.

project		
Likelihood of identification	High	A high rating is given as there are a lot of different variables including quasi-identifying variables (such as date of birth, address, indigenous status etc.) contained in the programs and payments data that will be used in the data integration project.
Technical complexity	Low	Technical complexity here refers to the output. That is, how difficult it is to confidentialise data for external publication and/or ensuring that external users who need access have access to unit record data. For example, external users may need access to linked data for research purposes. At this stage, external output is not proposed.
Pre-mitigation Likelihood Assessment	**Medium**	

This pre-mitigation likelihood assessment of medium aligns with the risk framework guidelines.

Overall pre-mitigation Assessment	**High**

Based on the above assessment, this project is classified as ‘high’ risk. The data is highly sensitive, with a large number of identifiable variables on both the programs and payments data. Having assessed the initial risk, DSS can now take actions to reduce the risk of undertaking this project.

4.1.2. Post-mitigation Assessment

There are a number of things that can be done to mitigate against the likelihood of a breach occurring. The data integration project can be designed in such a way that even if the consequence of a breach occurring is high, the likelihood is reduced such that the project does not require an accredited Integration Authority.

These are the mitigation strategies applied to reduce the consequence risk:

Elements	Reducing the likelihood of a breach occurring in the first place
Sensitivity	Initially, the data is assessed as highly sensitive. However, the project design is such that the entire dataset is not required. The separation principle plays a big role here. • It is proposed that a SLK would be created for programs and payments data using the same algorithm. This would negate the need for access to variables that are highly sensitive on the original dataset. • A file with the record identifier between the programs and payments data would be retained. This means that the linked file would not contain any sensitive data. • Only if a research request is approved would they be given access to the

	linked dataset. ○ DSS already has processes in place to handle research requests. ○ All internal research requests would also need to go through an approval stage.
Consent	The data being linked is all an administrative by-product. The purpose of this linking activity is to analyse client pathways through the whole social security system to improve programs and policies. Ultimately, this is the objective of the organisation and thus this project is a strategic move to enable us to provide sound policy advice and better design our programs to achieve quality outcomes.
Amount of information about a data provider	***Linkage stage:*** It is proposed that only five or fewer variables be used to create the SLK. These would include variables such as: • Name (surname and given name) • Sex • Date of Birth ***Storage stage:*** The linked file would only contain the SLK, weight (the strength indicator of the link) and record identifier. There is only one quasi-identifying variable in the linked file, as an SLK includes date of birth. This is not enough for identification. ***Analysis stage:*** Researchers (internal) would need to go through an approval stage and access would only be granted once it can be shown that the public benefit of the research outweigh the risk. Throughout the separation principle is applied.

Below are the mitigation strategies that DSS proposes to adopt to reduce the likelihood of a breach:

Dimension	**Initial rating**	**Mitigation strategies (reducing likelihood of breach)**	**Revised rating**
Managerial complexity	Low	DSS will be responsible for managing the internal stakeholders and ensuring that there is clarity around the complex data governance of this project. They will report to a steering committee to ensure that the risk of breaches are minimised. There would be clear terms of reference for the steering committee.	Low
Nature of access	Low	Different staff members require access to various aspects of the data at different stages of the project. To mitigate against this risk, the separation principle is applied throughout the project. ***Extraction stage:*** Staff members with appropriate security clearance will create the SLK based on the four variables identified above. ***File transfer stage:*** No external file transfer is required for this project.	Low

		Internally, access to the system/s would only be granted on a need-to-know basis. Internal data transfers (if required) would only be undertaken by staff with appropriate level of clearance. ***Linkage stage:*** Staff members responsible for data linking would only have access to variables needed for the linkage – in this case this is the SLK variable. ***Analysis stage:*** The internal researcher is only given data extracts required for their research. ***Storage stage:*** The variables from both datasets are never stored in full in a single file.	
Duration of the project	High	While the duration of the project is long-term, the stored linked file does not contain any variables that would pose a risk to individuals in the event of a breach. It's only when the content data is extracted for researchers that there is a risk of a breach. However, there are already policies and protocols in place (such as departmental protective security) to ensure this does not occur.	Medium
Likelihood of identification	High	Risk mitigation strategies can be applied to various stages in the data linking cycle, including extraction, file transfer, linkage, analysis and storage. ***Extraction stage:*** There are two extraction stages – one to create the SLK and the other to extract the content data for the researcher. In the first stage, only four variables are needed to create the SLK (and SLK in itself cannot spontaneously identify an individual). In the second stage, the extraction of content data for research purposes is already subject to systems that protect the privacy of data providers and confidentiality of data, including protective security measures. ***File transfer stage:*** As there is only one data custodian involved and access to data is already established, there are no security issues involved with file transfer. ***Linkage stage:***	Medium

		As SLKs are used to link the two datasets together, only one quasi-identifying variable (date of birth) used in the linkage stage. ***Analysis stage:*** The release of contents data to the researcher would go through already established practices and Protective Security. No name data is provided for analysis. ***Storage stage:*** The linked file would contain content variables, the SLK and a weight variable. There are no spontaneously identifying variables.	
Technical complexity	Low	At this stage, output is not proposed to be published externally.	Low
Post-mitigation Likelihood Assessment	**Low**		

Post-mitigation risk rating	**Medium**

In this particular case, an accredited Integrating Authority is not required[18].

Next Steps: Register project on the Data Integration project register. An accredited Integrating Authority is not required for this project.

[18] It is worth noting that an accredited Integrating Authority may also introduce an increased element of risk. Mitigation strategies would be required to address the introduced risk. For example, if an accredited Integrating Authority is required, then an extra file transfer stage is introduced. There would need to be mitigation strategies to minimise the introduced risk.