BALANCING THE PENDULUM

An analysis of data sharing and privacy

Author: Frazer Walker Partner Ian Chisholm*

frazerwalker

October 2021

Data is valuable. A host of organisations, including governments at all levels, insurance companies, banks, transport operators, health service providers and more, have massive datasets and many are happy to share de-identified data for the public good.

Data sharing can boost economic growth and propel innovation. But there are privacy and other ethical implications. Practical gatekeeper resources are required to monitor and manage data sharing.

There must be a balance between public benefit and privacy protection. The pendulum shifts, depending on the data involved and the depth of personal information it contains.

Frazer Walker works with many government and private sector organisations to implement data-sharing initiatives and assist them to put in place the guardrails to protect data and control against the risk of re-identification.

This white paper outlines the rapidly evolving field of data sharing and offers guidance on data-sharing methodologies for organisations to consider.

Questions that need to be answered before data sharing is enabled include – who wants access to the data and for what purposes? Where is the data stored and what security levels protect it?

While the framework 'Five Safes' is emerging as the foundation stone for data sharing, it has its critics. Risk management tools that aim to mitigate accidental data breaches and ensure re-identification cannot occur are important. Organisations working in the data-sharing space, like Frazer Walker, are continually seeking more advanced methodologies to protect personal information.

Explosive growth of data and datasets

The massive growth of data through the fourth industrial revolution (the ongoing automation of traditional manufacturing and industrial practices, using modern smart technology) has seen data become the 'new oil'. When that rich resource is sufficiently clean, it can be analysed for ethical purposes and produce insights that benefit the community.

When things go wrong, data can help to identify what occurred or find information that's not yet known. Organisations, including governments, not for profits and the corporate sector, are rapidly cataloguing their data assets and mining them for value – whether for commercial gain, service improvements, better customer experiences or greater value to citizens.

Governments are keen to unlock the value in their datasets for better citizen and societal outcomes and are moving to an 'open by default' posture, allowing de-identified and aggregated data to be released for further value extraction by third parties, including other government departments, universities, research institutions and not for profits, or even for commercial outcomes.

This can take the form of fully open datasets, that is, unrestricted access for whoever wants it, through to controlled data-sharing arrangements that require greater layers of defence, given the more sensitive data contained in the datasets.

De-identification key to protecting privacy

De-identification of large datasets is critical to protecting individuals' privacy and retaining the trust of customers and citizens whose personal data forms the basis of this type of information. However, the ability to blend datasets, whether private or public, risks individuals being re-identified.

A classic example of algorithms going wrong was the Public Transport Victoria (PTV) release of a Myki dataset in July 2018 containing 1.8 billion historical records of public transport users' activity. The Myki card, like the Opal card in NSW and the Go card in Queensland, is a reusable electronic card for travel on public transport.

The data was provided for use at the annual Melbourne Datathon, an event at which members of the Victorian data science community compete to find innovative uses for a dataset. The dataset contained the 'touch on' and 'touch off' activity records of 15.1 million Myki cards used over the three years to June 2018.

The dataset was de-identified before use, including being subjected to a privacy impact assessment (PIA), which assumed the dataset was successfully 'anonymised' by PTV and could be safely released for use.

During the datathon, some participants highlighted the risks of re-identification. This was subsequently achieved around September 2018 by academics from the University of Melbourne who could identify themselves and people known to them.

The dataset was revoked and CSIRO asked to investigate. CSIRO subsequently found the dataset's detailed nature created a high risk that some individuals may be re-identified through linking with other information sources.

The Office of the Victorian Information Commissioner (OVIC) published an investigation report [PDF 2.2MB] about the disclosure of the Myki data and highlighted the risks of the processes used. OVIC concluded that:

- De-identifying large and complex datasets is difficult and, in some cases, may be impossible
- Organisations should not rely solely on de-identification to protect data
- When considering whether 'de-identified' information is personal information, context is crucial
- PIAs, if conducted incorrectly, can create a false sense of security
- Clear lines of responsibility are needed for effective data governance.

The rise of the Five Safes framework

The Five Safes data-sharing framework was developed by the UK Office of National Statistics in 2002-03. The framework has since been used around the world, from the UK Data Service to the Australian Bureau of Statistics and the Australian Institute of Health and Welfare, specifically for:

- training people in how to manage sensitive data responsibly
- describing the management dimensions of sensitive data, and
- increasingly as a framework to implement good data control of sensitive and private data.

The Five Safes framework is designed to be used across all five dimensions to manage what type of sensitive data can be safely shared with other groups, departments and organisations. As such, there are different approaches to understanding and applying the framework when assessing the risks of sharing sensitive data with third parties and internal groups.

The five simple protocols aim to provide complete assurance for data owners when accessing confidential and/or sensitive data by ensuring the following 'safes':

- **Safe data:** data is treated to protect any confidentiality concerns
- **Safe projects:** research projects are approved by data owners for the public good, assessing each project on its public benefit outcomes
- **Safe people:** researchers are trained and authorised to use data safely; ethics and data handling training is a minimum for assessing people's readiness to handle sensitive data
- **Safe settings:** a secure lab environment prevents unauthorised use; tight controls over other dimensions of safe settings can be used in other circumstances
- **Safe outputs:** screened and approved outputs that are non-disclosive.

For example, 'safe settings' can be used to pass datasets to trusted third parties if their own safe settings are comprehensive, mature and well governed. However, if a third party's settings are not as robust as required, sharing data may involve securely accessing the data hosted in a protected lab via strong remote access services.

If the risk of re-identification is too high, safe settings may restrict access to very sensitive data by either method. For example, the UK Data Service Secure Lab provides access to sensitive or confidential data in a controlled manner for groups with approved researcher status, enabling researchers to access and use datasets in a secure, responsible way. The Five Safes are assessed together; the example above is for safe settings only.

Five Safes framework concerns

In 2020, three University of Melbourne academics wrote a paper, *Not fit for purpose: A critical analysis of the 'Five Safes'*, [PDF 251 KB] which raises issues with data-sharing risk management approaches, such as Five Safes, even challenging the idea that data sharing can ever be fully safe from re-identification. They suggest that governments must get the balance right between the interests of government in unlocking datasets based on personal information (the public interest test) versus the right to privacy.

Dr Chris Culnane, Associate Professor Benjamin Rubinstein and Professor David Watts argue that, despite its popularity, Five Safes has undergone little legal or technical critical analysis. They say Five Safes is fundamentally flawed: from being disconnected from existing legal protections and appropriation of notions of safety without providing any means to prefer strong technical measures, to viewing disclosure risk as static through time and not requiring repeat assessment.

"Five Safes provides little confidence that resulting data sharing is performed using 'safety' best practice or for purposes in service of public interest," they argue.

This type of research provides an opportunity to reflect on frameworks, tools and processes so far and to strengthen them going forward to ensure data is shared safely.

Frazer Walker's data-sharing insights

Frazer Walker has identified the following key discussion points that need further analysis.

DATA GOVERNANCE

Although there is a push to ensure frameworks for data sharing are encapsulated within algorithms, then automated to optimise speed to decision and reduce costs, judgement from data owners, stewards and custodians is still required on a case-by-case basis to manage and mitigate the risks of accidental data breaches or harm from unintended consequences.

ASSESSING SAFE SHARING

Additional tools are required for the more qualitative dimensions of the Five Safes framework – from a counterparty accreditation process through to data-sharing legal deeds, overseen by ethics committees, and good data governance from data owners, stewards and custodians of the datasets.

CONTROLS FOR RISKS

Clear controls for key data-sharing risks need to be established during the design process, enabling a basis for ongoing risk management and auditing across the three lines of defence – management control; risk control and compliance oversight functions established by management; and independent assurance.

ETHICAL RE-IDENTIFICATION

New tools always require significant amounts of testing to provide confidence in their suitability and reliability before implementation. Given the proliferation of external datasets that can be blended with key datasets, consideration may be required to extend the testing team's capabilities to include an ethical re-identification hack, similar to 'red teaming' for cyber security penetration tests. The purpose is to continually validate and verify that re-identification of key datasets is a negligible risk and provide confidence, particularly for legal and privacy requirements.

DATA-SHARING PLATFORMS

There is continued growth in data service providers that are fully accredited intermediaries for hosting key datasets that can be drawn on by agencies to assist in data sharing. They provide secure data-sharing platforms for data that is more sensitive to re-identification and therefore requires more controls than pure open government data.

PRIVACY IMPACT ASSESSMENTS

These may need to include letters of attestation from data custodians or stewards about the methods of de-identification and their suitability for the particular type of dataset as a basis for assessing privacy risks.

CROSS-DISCIPLINARY TEAMS

Data has no boundaries. Therefore, a team from multiple disciplines is best placed to establish data-sharing frameworks and tools to verify that data can be shared safely. Examples of capabilities include STEM personnel with backgrounds in data analysis, data engineering, data wrangling, analytics, data science and statistics, coupled with team members who understand the business context with humanities backgrounds and legal, privacy and risk capabilities. This will ensure multiple perspectives are considered at the design stage, which has a higher likelihood of mitigating risks further into the project.

THE DESIGN STAGE

As with the importance of putting the business process first, with technology meeting the business requirements, so too it's important that privacy and ethics are considered during the design stage. It is hard to retrofit privacy and ethics into processes and tools when they were not considered up front.

BUSINESS GLOSSARY

The increase in use and sharing of data with new tools and algorithms results in terms that require consistent, known definitions across all team members working on data-sharing projects. Examples are 'bias', 'harm' and 'unintended consequences'.

EXTENSION OF FIVE SAFES

If you're using Five Safes, build your data-sharing framework to integrate with new innovations, such as 'safe algorithms', in preparation for further automation that may use artificial intelligence (AI). 'Safe algorithms' could replace the 'safe people' dimension for processing data for analytical purposes (such as clustering or classification) or for delivering smart services (such as smart lighting or smart message routing). The 2018 ACS white paper, *Privacy in data sharing: A guide for business and government*, notes that the algorithms operate differently to human researchers and therefore "some of the implicit assumptions in the Five Safes framework need to be re-examined". Five Safes is a system model, intended to be considered in the context of all the elements. When a researcher (or algorithm) is permitted to access a dataset, there's an assumption all other necessary conditions are in place. "If secure facilities do not exist, this does not seem an appropriate way to use the data," ACS says.

DATA/AI ETHICS

Globally, there is a plethora of high-level principles to use as boundaries for the ethical use of data, algorithms, technology and AI. Sometimes the principles align with an organisation's existing governance, risk and compliance (GRC), or ethics, environment, sustainability and governance (EESG) disciplines. Baking these principles into the design stage of data sharing can mitigate against the potential misuse of data and AI.

Data sharing goes national

On 13 August 2021, data and digital ministers of all Australian jurisdictions agreed to a program of work for national data sharing. This followed the signing of the Intergovernmental Agreement on Data Sharing [PDF 700KB] by all Australian governments at National Cabinet on 9 July 2021.

The national data-sharing initiative prioritises data sharing and reform across three initial areas: natural hazards and emergency management; waste management; and road safety.

The ministers agreed future priority data-sharing areas could include family, domestic and sexual violence; closing the gap (a social justice campaign for Indigenous Australians); and veterans' health.

The data-sharing agreement recognises that data is a shared national asset and promotes its value and use to improve the lives of Australian citizens. As the agreement notes, "access to data is critical for policy, service delivery, and government decision making" and sharing data between governments is an efficient use of resources that will also "help drive economic value, innovation, improve services, and deliver better outcomes for Australians".

All jurisdictions party to the intergovernmental agreement view data sharing as the default position when it can be done securely, safely, lawfully and ethically, using established privacy standards.

The established privacy standards drawn on by the states, territories and Commonwealth have been defined in the 2019 document *Best practice guide to applying data sharing principles*, [PDF 700 KB] available on the Office of the National Data Commissioner website. The best practice guide draws on the Five Safes data-sharing framework.

Further underpinning the national data-sharing approach is the Commonwealth's new data legislation, the *Data Availability and Transparency Bill 2020* (DAT Bill), introduced to the Australian Parliament on 9 December 2020.

Although not yet enacted but expected before the end of 2021, having gone through several senate reviews, the Bill will help organisations to request controlled access to government data for three purposes:

- improving government service delivery
- informing government policy and programs
- research and development.

The Bill provides the legal framework for the Commonwealth to openly share data, based on the Five Safes framework. The DAT Bill will be overseen by the National Data Commissioner, an independent regulator that will ensure all scheme participants adhere to the scheme's safety and security requirements.

The Australian Prudential Regulation Authority (APRA) has produced several standards and guides that should be part of the toolkit for APRA-regulated entities to ensure compliance with data-sharing requirements. The key documents are:

- Prudential standard CPS 234 [PDF 268KB] Information security
- Prudential practice guide CPG 235 [PDF 328.44KB] Managing data risk

CPS 234 aims to ensure APRA-regulated entities take measures to be resilient against information security incidents (including cyber attacks) by maintaining information security capabilities commensurate with information security vulnerabilities and threats. A key objective is to minimise the likelihood and impact of information security incidents on the confidentiality, integrity or availability of information assets, including those managed by related parties or third parties.

CPG 235 is a prudential practice guide (PPG), providing guidance only, but likely to become a prudential standard. Frazer Walker has been working with clients conducting internal audits against CPG 235, establishing data governance functions, building data governance frameworks and record retention policies, and establishing data asset registers. The PPG is not all-encompassing but provides data risk management guidance on areas where weaknesses have been identified during APRA's ongoing supervision activities.

The path forward

If we seek the full value of data sharing across governments, private enterprise and not for profits, and understand the benefit to citizens for mobility, health, smart cities, and financial or government services, then we need to continue data sharing, ranging from open government datasets to controlled access to various levels of sensitive information with appropriately vetted and managed third parties.

There are myriad benefits of data-sharing frameworks. However, acknowledging the risks with any framework is wise and necessary. Developing a mature understanding of all the principles of Five Safes is critical to ensuring public trust and managing privacy risks while still delivering public benefit outcomes in a controlled, safe manner.

Frazer Walker encourages a robust debate about issues raised in this white paper to ensure the integrity of data-sharing principles.

*Ian Chisholm is a partner and co-owner of Frazer Walker. He has more than 25 years' operational and technology leadership experience in the insurance, banking and wealth management sectors. He has proven expertise in forming practical business strategies and building technology capability within organisations. His strengths lie in strategy and planning, governance and risk management, business process improvement, and information management.

Contact: ian.chisholm@frazerwalker.com; LinkedIn