Part 4

The Importance of Creating and Building an Evidence Base

—

Chapter

8

Maximising evidence-based policy analysis through data sharing

—

Professor A. Abigail Payne & Dr Rajeev Samarage

Each policy or decision a government makes costs money and will impact individuals and communities. Over this century data have increased exponentially and techniques for handling big data sets have improved. And yet, many data sets remain locked up, unavailable, or only provided to a select few individuals. This chapter provides the argument for leaping forward with the provision of data to trusted analysts so that we can build a more effective and timely evidence base to inform policy and practice in Australia. Yet, this chapter also stresses the importance of ***not*** being reckless or cavalier. We present a framework for both addressing the sensitivity and privacy issues when working with data and for ensuring shared collaboration in the development and structuring of research-ready data sets. The recent passage of the *Data Availability and Transparency Act 2022* (DATA) by the Commonwealth government provides the impetus for state governments and private data providers to adopt the principles that underpin DATA and for the rapid deployment of platforms to make greater use of data.

INTRODUCTION

With modern technological advances in computing, data management and storage, and business practices in the age of the internet, there has been an exponential rise in the amount of data we collect.[1] These data capture our interactions and behaviours and help us to understand better business and service provider finances and activities as well as community and household activities. Much of the information collected for purposes other than for research is, indeed, very useful for social science research. This is particularly true of the public sector with activities such as tax filings, health information, schooling attendance and performance, social security payments, and government expenditures being routinely collected and stored as part of business-as-usual activities.

Data access is a key barrier to better evaluation of existing policies and policy innovation

Gaining access to administrative data for analysis is often challenging and usually involves many layers of bureaucracy. All levels of government capture and hold data for administrative purposes. These data usually cover very large, if not all, of the relevant population. By capturing information on a relevant population, one can more easily study large-scale questions that pertain to that population. The large scale of the analysis is important, both in terms of minimising the risk of obtaining biased results that can happen when one works with a small and non-random sample of the population, as well as in increasing the power of the estimations from the statistical analysis. Moreover, by gaining access to administrative data, one can reduce the cost of research that would be associated with having to collect the information through other means, for example, through surveys. In many cases, the measures from administrative data are more accurate than the same measure if collected through surveys. Finally, because the measures are often collected repeatedly with administrative data, it is much easier to explore behaviour and outcomes over several periods, making it easier to discern patterns of behaviour as well as to address concerns in any analysis where not all measures relevant for the analysis are available for study.

In some parts of the world, such as the Nordic countries, administrative data have been available to researchers for many years (Connelly et al., 2016). In Australia, access to and use of administrative datasets remain under-utilised but use and access has been increasing in recent years due to recent legislation and advocacy and promotion by key stakeholders. Local, state, and Commonwealth government departments hold an extensive number of longitudinal administrative datasets but a lack of relevant frameworks and governance over these data have led to Australian researchers having to look elsewhere to obtain the data they need to study policy (Productivity Commission, 2010).

A 2010 Productivity Commission report concluded that access to de-identified administrative data for public sector staff and researchers be prioritised (Productivity Commission, 2010). In 2015, the Australian Government Public Data Policy Statement released by the Prime Minister mandated that the Australian government commits to optimising the use and re-use of public data to drive innovation (Turnbull, 2015). This use and re-use of data from the public sector (and private sectors too) can generate direct, indirect, and induced impact to data providers, data users and the wider economy respectively. The OECD (2019) shows that data access and sharing can help generate social and economic benefits of up to 1.5 percent of GDP in the case of public sector data and up to 4 percent when including private sector data. This notion cements data as an infrastructural resource and an investment.

4%

Benefit to GDP by enabling access to and sharing public and private sector data.

OECD (2019)

The real shift in thinking around data in Australia, however, came from the Productivity Commission's Data Availability and Use Inquiry in 2016 (Productivity Commission, 2017). This inquiry identified a 'lack of trust by both data custodians and users in existing data access processes and protections' and recommended the creation of data sharing and release policies that subsequently became the *Data Availability and Transparency Act 2022* (Cth). In addition to this legal framework, advances in technical frameworks for data access, such as that evidenced by the increased availability of secure infrastructure (trusted research environments), are leading a wave of hope of increased research access to administrative data from public sector sources amongst researchers. This is great progress for Australian data custodians, researchers and the community who stand to benefit from better policies.

[1] It is estimated that the amount of data in the world will grow from 44 zettabytes (ZB), 44 sextillion bytes, in 2020 (WEF, 2020) to 175 ZB by 2025 (International Data Corporation, 2018).

WHERE TO FROM HERE?

Secure and open data platforms that work together to house and transform data, as well as provide the ability to analyse the data by researchers, policy analysts, and service providers, will support cutting-edge research and ensure Australia plays a role on the international stage. While some platforms exist, there is scope for creating specialised platforms and for providing the mechanisms that support the virtual linking across platforms to enhance data use and to enable deeper and more rigorous analyses of policy relevant questions. These platforms should build on the work by the Office of the National Data Commissioner and the principles (and regulations) captured under the *Data Availability and Transparency Act 2022*. Given the diversity of disciplines and the range of approaches available for data creation, use, and analysis, plus political and governance issues associated with many datasets, a single platform will not suffice. To recognise the potential for using data to promote policy innovation, many platforms will be needed. Collaboration across platforms, as well as the importance of promoting research independence and following a high standard of protocols that permit transparent and verifiable status of the platforms, are necessary components for enabling effective use of data.

The power of administrative data, whether from public or private sources, is that there are many questions that can be studied using a range of domain and disciplinary expertise. Data not developed for research, such as administrative data, must be assessed for quality and transformed for the purposes of the type of analyses that will be undertaken. These assessments and transformations will vary across disciplines given the types of questions studied and the frameworks for studying these questions vary across disciplines. There is value, however, in encouraging the development of platforms that permit and encourage the sharing of knowledge, especially in the use and transformation of administrative data.

We are living in exciting times in that there is greater recognition today of the importance of providing access to critical data that can be used to understand, to improve, and to shape policy and practice, and in developing frameworks and protocols for the housing, transformation, and sharing of data. Equally important are the mechanisms for updating and developing processes for continual updating and improvement of the data assets used for research, analysis, and evaluation.

The purpose of this chapter is to focus on the importance of making better use of administrative data to inform, test, and shape economic and social policy. This chapter aims to provide an argument for why administrative data represents a game changing innovation for undertaking evidence-based policy analysis. It also aims to provide a framework to ensure the appropriate use and analyses of data. We also discuss how recent technological advances such as the use of trusted data environments are making it easier to apply the learnings from data analysis to policy. Finally, we provide an overview of the pitfalls when working with administrative data.

DATA ARE A GAMECHANGER FOR INFORMING AND SHAPING POLICY

The role of administrative data for undertaking social science research

Administrative data bring several opportunities to social science research. This is evidenced by the large rise in its use in economics research. Einav and Levin (2014) find that 26 percent of papers using any form of data published in the *American Economic Review* used administrative data in 2014 compared to 4 percent in 2006. While survey and field experiments remain important today, this chapter focuses on the need to use administrative data both as a tool in the toolkit but also as a means to complement and to enhance the value of survey and field study data. Before proceeding further, it is important to describe what we mean by 'administrative data'. Here we provide (right) a description of administrative data, its structure, and potential issues.

One of the most important advantages of administrative data is the **low cost** for use in research as the data have been collected for other purposes. In comparison, statistical surveys and experiments are expensive to design, develop and implement. While the setup cost for collecting, validating, and transforming administrative data may be high, the running costs are lower (United Nations, 2011). Another advantage of administrative data for social science research is the frequency at which data can be produced, mainly due to the reduced cost and reduced response burden to respondents and data suppliers. This also means that the data can be regularly updated, sometimes continuously, resulting in excellent sources of longitudinal data often of the same unit of observation, that is, individuals, households, businesses and so on.

A significant advantage of administrative data sources is that coverage is significantly larger, with sample sizes much greater than social surveys. Administrative data sources often provide complete or near complete coverage of the target population whereas sample surveys can often only directly cover a smaller sample. While social surveys implement techniques such as oversampling (sometimes referred to as 'boosting'), they still may not support in-depth statistical analysis of specific sub-populations as is possible with administrative data (Connelly et al., 2016). Administrative data are also particularly useful for studying issues where there is an absence of survey data. For example, administrative data offer the opportunity to create cohorts of individuals to study time-varying changes or to study the effects of significant events on individuals when there was no primary data collection at the time. Administrative data tend to capture a potentially more representative population as the data may capture information on those who may not respond to surveys—a feature that is particularly important in the study of poverty and disadvantage.

Here are a few examples of existing research papers that utilise administrative data to explore a range of economic issues.

Chetty et al. (2014) use administrative tax records in the United States to study intergenerational earnings mobility. Using cohort analyses of comparing earnings data from tax records for both parents and their children in adulthood, they conclude that young people entering the work force today have the same chances of moving up the income distribution as children born in the 1970s.

What are administrative data?

Administrative data are defined as data which are derived from the operation of administrative systems and processes. Unlike survey data and experimental data, these data are not collected for research or aimed to address well defined hypotheses. As often is the case, no research input was provided for the design, structure and content that is to be collected. Administrative data are usually very large with large numbers of observations and variables but depending on the organisation particular care needs to be applied to determine the actual sample captured within the data. If the administration collecting the data is a public sector agency as is often the case, this may not be much of an issue as nearly the entire population of individuals may be interacting with these systems. Administrative data are usually messy and often require significant data management capability such as cleaning, organisation and profiling before further analysis can be conducted.

Ananyev et al. (2022) use a 17-year panel dataset comprised of administrative tax records in Australia to better understand who experiences major earnings shock during their working life and what we know about recovery of these shocks. This study permitted an understanding of how changes in the macro-economy affect shocks and recovery, how shocks and recovery vary across age groups, and how changes in family composition (getting married, having children) are correlated with these shocks and recoveries.

Deutscher (2020) also uses Australian tax data to measure the causal effect of neighbourhood location when growing up on adult income and other socio-demographic measures.

Zajac et al. (2021) use what is known as the Multi-Agency Data Integration Project (MADIP) from the Commonwealth government to explore labour market earnings of recent university graduates. They observe differences in earnings and earning trajectories based on the socio-economic background of the students studied.

With greater access comes greater insights

Access to administrative data is only the first step. These data could be instrumental in gaining better insights into a range of issues that affect Australians. Our society today has evolved where most economic and social policy challenges are complex. The complexity of these issues means that to undertake analyses to support policy innovation requires information that reflects many facets of a person, household or community. Below we provide two examples to illustrate how administrative data can support greater complex analyses.

EXAMPLE 1: UNDERSTANDING SCHOOL LEAVERS

Why do some students leave school before completing Year 10? In Australia, using 2016 Census data, analysis by Marchand and Payne (2022, Chapter 7 of this Compendium) illustrates that by the age of 24, close to 20 percent of the population has not completed Year 10. A typical student would complete Year 10 by the age of 16. Since 2010, the National Youth Participation Requirement expects that all youth will participate in schooling until they complete Year 10. What can we do to promote the achievement of this goal? Before we can answer this question, one would want to understand why students may leave school before Year 10.

With administrative data, we can collect information specific to the student, such as attendance rates, grade progression, and performance in school (on standardised tests and in the classroom). We can also capture information about the student's household, which can range from a better understanding of the family dynamics, including moving locations, household size, and other characteristics that can impact the living environment of the student. Further, we can capture information about the student's schooling environment and residential community. By observing the student and their environment over time, we can better assess the core factors that might impact the likelihood of leaving school early. And with close to the population of all students, we can better understand differences across specific geographies, the impact of service provision to support students who might be at risk for leaving school before year 10 and consider targeted approaches for achieving the goals of the National Youth Participation Requirements. By capturing geographies across Australia, we can better understand why interventions work in some areas but not others, allowing for both a general understanding of the issues associated with school completion as well as an exploration of how best to provide tailored but tested solutions across geographies.

EXAMPLE 2: BREAKING THE CYCLE OF DISADVANTAGE AND REDUCING INCOME POVERTY

Payne and Samarage (2020) and Ananyev et al. (2020) document that overall poverty rates in Australia have been relatively stagnant over the last decade and that there are many communities where poverty rates are alarmingly high. They also show that education and employment are highly correlated with exiting poverty. Vera-Toscano and Wilkins (2020, 2022) document that a young adult who experiences economic disadvantage as a child is more likely to experience poverty as an adult than a child who experiences no economic disadvantage growing up.

Encountering economic disadvantage and/or falling into poverty, however, is not simply tied to education and employment. As eloquently illustrated by Mallett and Cooney-O'Donogue (2019), income poverty can be accompanied by housing or food insecurity, being socially disconnected, not having good information about where to get help or what services are available, and much more. There are many economic and social drivers that can lead to falling into poverty, exacerbate one's circumstances or limit the ability to exit from poverty. There usually is no single silver bullet that will drive a big reduction in or prevent poverty. But there are multiple solutions that can support the alleviation of poverty. Which solutions will work best will depend on the family circumstances, the economic conditions, and the community and social context of the areas in which a given family lives. Not too surprisingly, research illustrates that to tackle poverty we should look at the combination of economic, social, and psychological factors and how they relate to each other (see, for example, Bossuro et al., 2022).

The complexity of the factors that contribute to poverty as well as the opportunities and support that can be provided to support an exit out of poverty and/or to prevent an entry into poverty is an area of work that would benefit greatly from increased access to administrative data. Data at the individual or household level, as well as community-based measures, are available from many publicly run sources and can even include privately sourced data. Accessing data about education, employment, housing, social welfare, tax, and interactions (positive and negative) with a range of departments or service providers helps to address the complexity of the issues and to provide more targeted insights into the possible ways to address poverty as well as to provide better guidance on the factors that can contribute to a set of interventions being more successful than a second set of interventions.

The power of administrative data to complement and inform survey data collection and field studies

Administrative data provide a necessary and complementary component to other methods for collecting data and undertaking analyses. In many instances, the data can provide insights into policy issues on their own. Sometimes, however, the data can be complemented with publicly available information that can be coded into a dataset and added to the administrative data. For example, data on matters that relate to activities in a community such as plant opening/closings, bush fires/droughts/floods and pandemic lockdowns, as well as changes in policies and other factors, can assist in supporting analyses by including explanatory information that affects one set of communities but not another set of communities. These additional measures might also provide a context for exploring the effects of an event on a community or a given population.

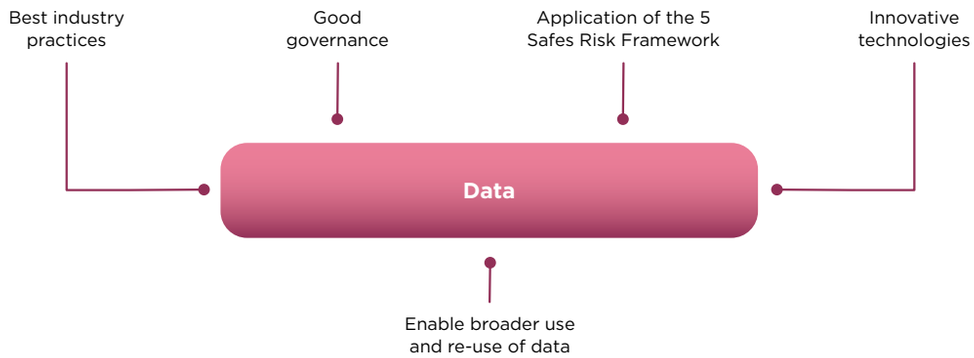
Administrative data, however, can also be used to enhance datasets that are generated through surveys. There are at least three ways in which administrative data can be used to support survey data collection and/or analysis. First, administrative data can be used to identify trends in behaviour, geographic areas of interest, and similar conditions. By undertaking analysis before survey development, the questions asked in the survey and the types of respondents pursued for the survey will be more refined than if one starts with a hypothesis and general knowledge of an issue that has not been tested. Second, if the survey dataset contains geographic and other demographic identifiers, then administrative data can be used to capture information about the geography (or population) under study that is not collected as part of the survey. Third, if permission is gained to link survey responses to administrative data, then the questions asked on the survey can focus more on matters that are not easily captured through administrative data. For example, in understanding poverty we might want to better understand the role played by being socially disconnected, the efforts undertaken to exit from poverty (for example, the types of jobs one has pursued, success in gaining interviews), as well as knowledge of the services available to support one's situation. Similarly, with survey data that are linked to administrative data, it is possible to make use of the survey data to explore backward-and forward-looking questions.

The opportunities for creating better field studies and randomised control trials would be increased if one had access to administrative data with measures that could affect decisions on the population to study and/or the types of experiments to run as part of the trial. Beyond the trial, administrative data can be used (if linked to the respondent and/or at least available with a relevant geographic identifier) to explore longer-term effects of interventions introduced through the trials. Administrative data could also be used to follow subjects who drop out of studies to better assess the potential biases introduced as a result of attrition in the original sample.

FRAMEWORK FOR ENSURING APPROPRIATE AND BROAD USE OF DATA

In this section we focus on providing a framework for maximising the value of administrative data for social science research. This framework focuses on: (1) appropriate governance; (2) the use of best practices for creating research-ready datasets; (3) the use of trusted research environments; and (4) the use of innovative technologies and practices from other research domains. Figure 1 outlines the framework.

Figure 1.
Framework for ensuring appropriate and broad use of data.



Good governance around ethical use of data

Administrative data often contain personal and sensitive information. Before any dataset is used, one must follow a set of practices that includes: (1) a consideration of ethical issues for using the data; (2) a structured approach that allows for the transformation of the data in a manner that permits the de-identification of records; (3) a governance and access process that minimises the risk of identification and/or release of the personal or sensitive information in a manner that would be harmful; (4) a consideration of where best to house data and the processes followed for accessing the data using a secure and protected environment; and (5) a development of practices and procedures for vetting analyses outside of any secure environment to ensure the information that is released cannot be used to re-identify the individuals used in the analysis. These practices illustrate the importance of developing frameworks to enable wider access to administrative data sources. These frameworks should cover multiple dimensions: legal; policy; organisation; and technical (United Nations, 2011).

Use of best practices

One of the best research practices to ensure the efficient use and re-use of administrative data is ensuring appropriate data quality. Administrative data can be messy and inaccurate. Information can be inputted incorrectly or not filled in at all (creating missing information). Moreover, the specific information collected and/or the process used for collecting the information can change over time. These changes can result in inconsistencies in the measures created. Thus, before any dataset is accessed, an assessment of the quality and consistency of the information collected is required. Failure to assess the data quality and understand the potential effects of data quality can lead to significant biases in the output. A crude but effective example of this could be an operational process where operators enter a zero to an income field within an organisational database when respondents have not provided a response. Failure to treat this variable to remove the zeros would lead to bias in the income distribution in the data used for analyses. As such, researchers often spend a large amount of time cleaning and understanding the data before the research analysis can commence. In some fields, such as data science, it is normal for researchers to spend up to 50 percent of their work time on cleaning and organising data (Anaconda, 2020) with some cases going as high as 80 percent (Lohr, 2014).

In instances where researchers use the same data asset across different institutions or sub-units within an organisation (or sometimes within the unit itself), researchers often perform the same processes to create a 'research-ready' dataset with no communication or sharing of information with other users of the same data.

There is currently no common understanding of what is defined as a 'research-ready dataset'. Work done by McGrath-Lone et al. (2022) involves a thematic analysis of relevant publications to define five broad characteristics of a research-ready dataset. These are data usability for research ('enhanced'), data accessibility ('access'), data comprehensiveness and ability to link to other data ('broad'), data transformation and quality checks ('curated'), and data documentation ('documentation). Our experience and that collective captured in the 60 years of working with data at the Melbourne Institute, and more recently through data curation activities within the Melbourne Institute Data Lab, informs us that additional practices around data reproducibility and data profiling should be included when creating a research-ready dataset. Hence our description (right) of a research-ready dataset and its features that have positive impacts on social science research practice.

What is a 'research-ready' data set?

A 'research-ready' data set is a data set that has undergone a range of technical tasks such as data transformation, harmonisation, data cleaning and preparation; as well as standardisation and documentation. The aim of creating such a data set is to do sufficient processing of the data to reduce the technical burden on the researcher and make the data available for **broad research purposes**. Subsequently, more processing could be performed by experienced researchers in particular sub-themes of research to create an 'analysis-ready' data set that is aimed at **investigating specific research questions**.

A research-ready data asset should have the following features to be useful and have a meaningful impact on social science research practices:

• The data should be meaningful and usable for research and statistical purposes including standardisation to an appropriate format.

• The data should be accessible with relevant permissions to 're-use' data for research, and appropriate frameworks in place (such as the Five Safes frameworks) when determining access.

• The data should have sufficient robustness checks and data quality checks to ensure data transformations are accurate and correct.

• The data should be transformed to increase research potential such as through harmonisation or linked to a range of other data sources (using a linking key at a unit level or using another appropriate variable) to increase the value of the data.

• The data should include indicators or measures of relevance and new derived variables relevant to research.

• The data should be accompanied with information and documentation about all aspects of the data including data collection, sampling framework and representativeness, data profiling to understand the extent of missingness and its impact on the captured sample, data quality statements and information on variables and derived variables.

• The data should be accompanied by relevant programs, code or scripts that enable users to replicate its creation from the 'as provided' data by the data custodians. These programs and code should follow best programming and coding practices including appropriate naming of variables, use of appropriate programming constructs to minimise repetition of code and the use of appropriate comments and descriptions where applicable.

DATA PROFILING TO UNDERSTAND THE DATA

Having a standardised definition for a research-ready dataset enables researchers to have a sequence of steps to create research-ready data assets that maximise value for research.

These steps, highlighted in Table 1, start with the formulation of the research question for which the data will be used and proceed to data profiling. Data profiling is a key step to interrogate the data using a range of descriptive analyses to understand its sample, data quality and missingness. Understanding missingness, the occurrence of missing values in the data, is crucial as it directly impacts the sample captured in the data. Missing data may be random (no systematic differences between missing data and complete data) or not random, where subsequent analysis may be biased if not handled correctly. For example, if you are using administrative tax return data to study people entering into poverty, removal of non-complete longitudinal histories for people with low incomes would be detrimental to the overall research design.

Another crucial aspect of the second step is to open a dialogue with the data provider to understand other details that may not be fully captured in the data documentation. This includes data collection protocols and any changes to these during the time that data were captured, treatment of variables, data quality statements and so on. The final and most critical stage is ensuring that the whole process is documented through memos, analysis plans and analysis code (see below) that allows the researcher to replicate the process used to take the data 'as provided' by the data provider to the research-ready dataset.

INCREASING REPRODUCIBILITY OF THE DATA

Once the data have been cleaned and they have been made ready for research, there is often the need to replicate this process of data cleaning. Sometimes this may be to reproduce another version of the research-ready dataset with variations required for specific themes of research. Often this is simply another form of documentation to ensure efficient re-use of the data by someone who was not involved in the cleaning and transformation processes. This replication activity is usually done by sharing of analytical code, scripts or programs that replay a set of instructions that take the 'as provided' data from the data custodian to create the research-ready version of that same dataset. This is a practice also increasingly followed by academic journals where researchers publish their data and code to increase scientific transparency, reproducibility and re-use of data. However, in practice, it is often easier to write new code than reuse old code. A recent study by Trisovic et al. (2022), which analysed over 9,000 unique R files to reproduce over 2,000 replication datasets, found that three of four files failed to complete without error on the first try. A little over half of the files still failed to complete after code cleaning techniques were applied. This highlights that there is a need for knowledge and practices that are commonly used in other fields such as the computer sciences to be effectively translated to research settings.

Table 1.
Framework for creating a 'research-ready' dataset.

Formulate the question	The first step is to formulate the research theme and/or research questions for which these data would be used. These questions need to be broad as 'research-ready' data are intended for broad themes of research while 'analysis-ready' data are intended for focused questions within the research theme. Formulation of this theme will aid in focusing on the types of data or specific measurements required for research. This also aids in identifying the data assets that need to be sought out to support this research.
Understand the data	A very critical step is to understand the data. Some key questions that should be asked and answered are as follows. • What measurements are captured? • How were the data collected? • How is the sample defined? • What additional measures are required or useful to collect? • What is the unit of observation? • Are there missing observations and/or measures? • Are measures consistently collected over time and/or are measures added or dropped over time? • How does having missing information affect my sample? If we have repeated observations for the same individual (or relevant unit), do we observe consistent information for the non-time-varying information (for example, gender, birthdate), and can we make sense of the time-varying information or fill in missing information? Administrative data capture usually follows administrative procedures that may not place a high weight on data quality captured. For example, measures such as address may not be updated. This step is about opening a dialogue between the research team and the data provider to ensure all information relating to data collection (and any changes to procedures of data collection) are identified.
Process the data	Once a deep understanding of the dataset as it currently stands has been identified, the next step is to begin processing the data to create the research-ready data asset. The data may need to be transformed to a specific format, and data cleaning, imputation and derivation of new variables will need to be done to support analyses of the research question(s) identified earlier.
Documentation	While this step appears here in this table, it is a step that must start when starting step no. 1 above. All decisions and aspects of research-ready data creation must be documented to maximise their re-use by other researchers. Documentation also includes the code, programs and scripts used for its creation. Appropriate coding standards and practices, which include naming conventions, proper commenting practice and use of complex code structures to minimise code repetition, need to be followed.

Use of trusted research environments

One of three reasons provided in a Productivity Commission report (2013) on why Australia lacks a culture of information sharing and data release is the protection of privacy. The other two are lack of data quality and concerns by governments about adverse findings on policy effectiveness. The security and confidentiality of individuals and businesses that interact with government systems is key to ensuring the public's trust in government to handle their data and subsequently use them for informing future policy. Opening data access to parties outside government, and in some cases other departments of government, increases the risks of disclosure.[2] In recent years, data providers have been more open to providing access to parties outside of government through the use of risk frameworks. One framework used in Australia is the Five Safes framework (Ritchie, 2008) used by several Australian government agencies, including the Australian Bureau of Statistics, as well as national statistical organisations overseas.

The Five Safes framework is applied across five distinct domains: projects, people, data, settings (what is determined to be the infrastructure used for data access) and outputs (from this infrastructure). These domains are usually applied in the following ways:

- **Safe Projects** Data applications undergo screening to ensure project goals and research purposes are aligned with the authorised purposes for data use.
- **Safe People** People who wish to access the sensitive data undergo appropriate vetting and training prior to being granted access.
- **Safe Data** Confidentiality of data units (that is, individuals, households, businesses, etc.) are protected through the use of statistical techniques such as de-identification, suppression, aggregation, top/bottom coding, random noise, etc.
- **Safe Settings** Security settings and controls in place for the environment used for data access and use are assessed to ensure they are appropriate.
- **Safe Outputs** Strict controls are enforced around which data (if any) and outputs derived from the data can be taken out of the safe setting (above). This includes the application of statistical disclosure control techniques to vet outputs before they are released to researchers.

One of the main drivers enabling data access is the security of the environment (safe settings) used for data access and use. Safe settings, together with the application of controls to minimise disclosure risk through the other four domains, ensure that researchers can access sensitive data in a way that no longer greatly limits the type or detail of data that researchers can access. Traditionally some of the requirements for these settings were achieved using physical purpose-built rooms where researchers used to access specific data and were unable to take any data outside these rooms. Today with advances in network infrastructure and cloud computing, a range of information security controls can be imposed on researcher environments and tailored at a user or project level.

There remain multiple instances, however, where data providers securely transfer and extract the data to individual researchers who are then responsible for their storage, analyses and destruction. But this responsibility is based on 'trust' and does not implement a safe environment framework. Such practices are costly and inefficient both to data custodians, as there is often a repetition of common tasks, and to researchers, who must adhere to technical and governance barriers imposed by data custodians. The secure and shared data environment model is a centralised approach that helps reduce these costs and inefficiencies.

In Australia, over the years, agencies that conduct data integration, Accredited Integration Authorities (or Accredited Data Service Providers under the new scheme outlined by the *Data Availability and Transparency Act 2022* (Cth), conduct high-risk data integration activities using a range of data sources and have used secure data environments to provide access to researchers and policy-makers. The ABS DataLab is one example of a secure data environment run by the Australian government. There is an increasing use of secure data environments to broaden access to detailed and sensitive data in a safe and secure way (Department of the Prime Minister and Cabinet, 2021), especially from non-governmental organisations. One such environment is the Melbourne Institute Data Lab (MIDL), a protected-level environment that enables the housing, curation and analysis of sensitive data. This is an investment led by the Melbourne Institute because of the importance of being able to create 'shared' environments that allow authorised researchers to access information from different sources, additional information contained in research-ready data assets enabling faster and better analyses of data. These shared environments also provide the unique ability to bring in private sector and public sector data with the aim to answer multiple research questions under a singular research/policy theme.

At varying levels of security settings and certifications these environments provide a good coverage for a range of security requirements imposed by data custodians. But more work needs to be done in this space to move away from the 'one size fits all' mode of thinking to understanding that multiple secure environments provide an excellent platform to data custodians to share their data with a large group of researchers. Moreover, there is also a need for better communication and integration between these systems. The current infrastructure landscape for social science research is fragmented across multiple systems with different resources, user bases, standards and protocols. There is a gap in how these systems can communicate with each other and share sensitive data, which makes it easier for researchers to migrate systems to better suit their research needs. This is the primary focus of the Integrated Research Infrastructure for the Social Sciences (IRISS) project led by the Australian National University with participation by the Melbourne Institute, the Institute for Social Science Research at the University of Queensland and the Australian Urban Research Infrastructure Network.

[2] Disclosure, or breach of confidentiality, of an individual, a group of individuals or a business being re-identified through a data release.

Use of innovative tools

THE USE OF APPLICATION PROGRAM INTERFACES FOR DATA ACCESS

One technological feature that is seeing increasing use for data access is via application program interfaces, or APIs. APIs are commonplace in computer sciences, specifically in fields of web development for websites or social media sites to request and share data from other web sites. In social science research there is often a need to access data stored in a web page or a series of web pages. This could include names and locations of child-care services, information relating to COVID-19 policies as they came into effect, obtaining location data for businesses in a business register or accessing labour force statistics for a specific area in Australia. Traditionally this would have been done manually or through the use of web scraping, the process of using automated scripts to extract information and content from a web page. Nowadays, the use of APIs allows researchers to request the data they want from the website directly through programming languages of their choice and plug the result directly into their research analyses. This is an amazing achievement for data sharing in the social sciences, but more work needs to be done to better inform researchers about how to access and utilise these tools for research. Another avenue for APIs is enabling data re-use. At the completion of a project or report, researchers could provide an API that allows other researchers to request relevant data stored on a web server or secure cloud repository. This would be effective in cases where the dataset is complex and unstructured and quite large in size.

INTERACTIVE DATA VISUALISATION AS A TOOL TO BETTER ENGAGE POLICY-MAKERS

Over the last three decades data visualisation has made a significant impact on how we explore data and extract insights. Given the increasing complexity of data, as is the case with administrative data, data visualisations offer a more user-friendly way to communicate key findings or insights with a range of stakeholders. Its visual and engaging nature also caters to communicating with stakeholders such as policy-makers who may not hold specialist expertise to interpret a table, for example, one with coefficients from a regression model. Nowadays data visualisation has expanded further into multiple focus areas through the use of interactive dashboards, data discovery tools and interactive visualisations for visual 'story telling' of data-derived insights. These technologies have also been adapted to sharing insights on social science issues by news agencies,[3,4] research organisations (Taylor, 2014),[5,6] and government agencies.[7] Interactive visualisation tools are engaging and can transform data into different visualisations with the click of a button. These tools make it easier for non-specialists to interact with the data and glean insights and trends in the underlying data at their own pace or interest. Due to privacy constraints, most visualisations use publicly available data or 'safe data' that have been vetted out of secure data environments. Most of the time this would mean data used are aggregated to some extent, limiting the level of data exploration possible with the flick of a finger. There needs to be more work in how technical controls can be used to make possible interactive visualisations that work for data held in secure data environments without impacting privacy and confidentiality. These methods can ensure we can move from data visualisations that are used to inform on today's story, to visualisations that are regularly updated to evolve with new incoming data as they are available.

3 'Income mobility charts for girls, Asian-Americans and other groups. Or make your own', New York Times, 2018

4 'Australian Census Explorer', SBS, 2022

5 "Debt in America, An Interactive Map", The Urban Institute, 2022 <https://apps.urban.org/features/debt-interactive-map/?type=overall&variable=totcoll>

6 "Breaking Down Barriers Data Visualisations", Melbourne Institute: Applied Economic & Social Research, 2022 <https://melbourneinstitute.unimelb.edu.au/research/reports/breaking-down-barriers/data-visualisations>

7 "Specialist homelessness services annual report, 2019-20", Australian Institute of Health and Welfare, 2020 < https://www.aihw.gov.au/reports/homelessness-services/specialist-homelessness-services-annual-report/interactive-data visualisation>

CHALLENGES WHEN WORKING WITH ADMINISTRATIVE DATA

Issues with access

One of the first issues with the use of administrative data for research purposes is ethics. Administrative data are not primarily collected for research purposes and, as a result, the public may have issues over their use and linkage to other administrative data sources. Currently this is managed through frameworks such as the Five Safes framework that ensures data are de-identified before release, the use of additional training and vetting before granting access, accessing data from a secure setting where access is controlled and monitored and having various output vetting processes in place. However, due to this nature of the data, access is still an issue. This is the case when requiring access to data from the Australian government while not in Australia. In most secure data environments, access to projects by existing, already authorised users, is not possible. This is particularly problematic in cases where academics are travelling or for affiliated academics overseas who want to study Australian economic and social policies.

Other issues with data access relate to the time and effort required to get access to administrative data. Time spent applying for access and undertaking training limit the time researchers have to do exploratory analysis. Access through secure data environments also greatly limits the possibilities for other researchers who do not have access to the same data to replicate results. Connelly et al. (2016) argue that making data analysis code and documentation accessible would allow researchers to examine research practices and build on existing work in the future.

Issues with data quality

As noted above, administrative data are often less systematic and require more data management to create research-ready data assets in contrast to traditional types of social science data. Administrative data may have large numbers of repeated observations based on how individuals or businesses interact with the administrative organisation. Administrative data may also have a large number of missing values based on the data collection procedures in place at the time. There may also be limited or no information about how data collection procedures and staffing change over time. Administrative systems may have variations in these aspects unlike that found with primary collected data where a team of interviewers undergo training and perform the data collection in a specific time-frame. These differences ultimately lead to lower data quality in contrast to primary collected data. It is generally accepted that data captured through surveys have a range of sources of errors, including measurement error, processing error, coverage error, sampling error and nonresponse error (Groves and Lyberg, 2010). Groen (2012) shows that administrative data may also contain measurement error arising mainly due to reporting differences. For example, the value that a respondent writes down as income for a social survey is different from that which they provide for tax purposes. While the latter seems more accurate, the reporting difference arises from the timing of the data collection. The survey may be collecting data monthly while the tax office collects data annually resulting in a seam effect where variations are larger in the administrative data over the survey data.

There are instances where data providers see the value in the data they capture for research purposes. In these instances, they often invest internally to curate and improve the data quality prior to release to researchers. With increasing budget and time constraints, however, the line item for data production for research purposes, especially when it is outside the remit of the organisation, may be the first to go. There needs to be a significant contribution of time and resources, including analysts with specialist technical expertise, to understanding, cleaning and curating data for research use. Often these data skills are not taught to social scientists in any depth, but a range of programs are available for data scientists. As Einav and Levin (2013) point out, often when companies are hiring for 'data scientists' to undertake such activities, they are generally looking for people trained in computer science rather than econometrics. For future economists who wish to work with large-scale administrative datasets, it is recommended to acquire some new skills in tools used by computer scientists, such as R, Python and SQL for conducting analyses, as well as some specific skills in efficient memory management such as compression, chunking and indexing.
A key takeaway for budding social scientists (and data scientists) when working with administrative data should be to learn the concepts of data preparation but understand and assess the effects to the sample when conducting these steps. Administrative data have large numbers of observations making statistical significance less relevant, but economic theory needs to be applied to formulate the research question, hypotheses and to apply reasoning for understanding attitudes and behaviours.

Issues with documentation

There is often a lack of clear documentation accompanying administrative data. As administrative data are often captured through business-as-usual activity, there are no clear descriptions, data quality statements, and metadata that form the underlying data documentation. This is not the case with primary data sources where there is a wealth of data documentation provided by the data collector and the data production team. Data collection processes, including questionnaires in the case of surveys, are described in great detail for social scientists to understand the implications of the measures that are captured. With administrative data, there is often limited or no documentation on the data collection/generation process. Additional effort is required by researchers to understand this process and its implications for the measures captured by the data. Often with administrative data, researchers begin projects without full knowledge of the data. This is the main reason why investment in creating research-ready data and documentation for re-use, and enabling the sharing of data, analytical code, and documentation with other researchers through shared data environments, will have dramatic returns for social science research practices in the future.

CONCLUSIONS

From a policy evaluation and innovation perspective, we should increasingly demand the building of a relevant evidence base to inform and test ideas. Like many countries, Australia has public and private data holdings that could be used to achieve deeper and more rigorous insights.

Critical to the building of the evidence base are the following components:

- Making it easier to access data by trusted users, such as has been articulated in the *Data Availability and Transparency Act 2022* (Cth).
- Encouraging best practices to understand, document and transform administrative data into research-ready datasets.
- Promoting and enabling stronger collaboration across research teams and organisations, especially to enable easier data access and better development of research-ready datasets.
- Building an eco-system of trusted data platforms where each platform provides unique and important capabilities given the diversity of disciplines and issues for which the same datasets can be used but also permitting the platforms to coordinate and collaborate with each other.
- Ensuring relevant infrastructures are in place to provide a high standard of expectations that relate to data governance, legal agreements and data access and use to address concerns that relate specifically to the sensitivities and ownership of the data that can be used for research and policy analysis.

We have seen over the past decade many inroads for better and more effective use of data and the building of evidence to help us to address the complex issues faced by Australia (and the world). But much more progress is needed.

REFERENCES

Anaconda (2020) '2020 State of Data Science: Moving from Hype towards Maturity', Anaconda Inc. https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf

Ananyev, M., Payne, A.A. and Samarage, R. (2020) 'Measuring Individual Poverty', Breaking Down Barriers Report #3, Melbourne Institute: Applied Economic & Social Research, Melbourne.

Ananyev, M., Payne, A.A., Wilkins, R. and Zilio, F. (2022). 'Prevalence and Recovery From from Negative Earnings Shocks: Evidence from Three Decades of Longitudinal Tax Data', *Breaking Down Barriers Report*, Melbourne Institute: Applied Economic & Social Research, Melbourne.

Bossuro, T., Goldstein, M., Bassirou K., Karlan, D., Kazianga, H., Pariente, W., Premand, P., Thomas, C., Udry, C., Vaillant, J. and Wright, K.A. (2022) 'Tackling Psychosocial and Capital Constraints to Alleviate Poverty,' *Nature*, vol. 605, pp. 291–297.

Chetty, R., Hendren, N., Kline, P., Saez, E. and Turner, N. (2014) 'Is the United States Still the Land of Opportunity? Recent Trends in Intergenerational Mobility', *American Economic Review Papers and Proceedings*, vol. 104–5, pp. 141–47.

Connelly, R., Plaford, C.J., Gayle, V. and Dibben, C. (2016) 'The Role of Administrative Data in the Big Data Revolution in Social Science Research', Social Science Research, vol. 59, pp. 1–12.

Department of the Prime Minister and Cabinet (2021) *"Australian Data Strategy: The Australian Government's Whole-of-economy Vision for Data"*, Department of Prime Minister and Cabinet, Commonwealth of Australia, Canberra.

Deutscher, N. (2020) 'Place, Peers, and the Teenage Years: Long-Run Neighborhood Effects in Australia,' *American Economic Journal: Applied Economics*, vol. 12, no. 2, pp. 220–249.

Einav, L. and Levin, J. (2013) 'The Data Revolution and Economic Analysis', Technical Report, NBER Innovation Policy and the Economy Conference, Washington, DC, NBER working paper 19035.

Einav, L. and Levin, J. (2014) 'Economics in the Age of Big Data', *Science*, vol. 346, 6210.

Groen, J.A. (2012) 'Sources of Error in Survey and Administrative Data: The Importance Importance of Reporting Procedures', *Journal of Official Statistics*, vol. 28, no. 20, pp. 173–98.

Groves, R.M. and Lyberg, L. (2010) 'Total Survey Error: Past, Present and Future', *Public Opinion Quarterly*, vol. 74, no. 5, pp. 849–79.

International Data Corporation (2018) 'Data Age 2025: The Digitization of the World from Edge to Core', IDC White Paper #US44413318, International Data Corporation, Massachusetts. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Lohr, S. (2014) 'For Big-Data Scientists, "Janitor Work" is Key Hurdle to Insights', *New York Times*.

Mallett, S. and Cooney-O'Donoghue, D. (2019) 'The Experience of Poverty, Then and Now', in Saunders, P. (ed.) *Revisiting Henderson: Poverty, Social Security and Basic Income*, Melbourne University Press, Melbourne.

McGrath-Lone, L., Jay, M.A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wiljaars, L. and Gilbert, R. (2022) 'What Makes Administrative Data "Research-ready"? A Systematic Review and Thematic Analysis of Published Literature', I*nternational Journal of Population Data Science*, vol. 7, pp. 1–6.

Meyer, B.D. and Mittag, N. (2019) 'Combining Administrative and Survey Data to Improve Income Measurement', NBER working paper 25738.

OECD (2019) *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use Across Societies*, OECD Publishing, Paris. https://www.oecdilibrary.org/sites/276aaca8-en/index.html?itemId=/content/publication/276aaca8-en

Payne, A.A. and Samarage, R. (2020) 'Spatial and Community Dimensions of Income Poverty', *Breaking Down Barriers Report #2*, Melbourne Institute: Applied Economic & Social Research, Melbourne.

Productivity Commission (2010) *Annual Report 2009–10*, Annual Report Series, Productivity Commission, Canberra.

Productivity Commission (2013) *Annual Report 2012–13*, Annual Report Series, Productivity Commission, Canberra.

Productivity Commission (2017) *Data Availability and Use*, Report No. 82, Productivity Commission, Canberra.

Ritchie, F. (2008) 'Secure Access to Confidential Microdata: Four Years of the Virtual Microdata Laboratory', *Economic and Labour Market Statistics*, vol. 2, no. 5, pp. 29–34.

Taylor, P. (2014) *The Next America: Boomers, Millenials and the Looming Generational Showdown*, Pew Research Centre, Washington, DC.

Trisovic, A., Lau, M.K., Pasquier, T. and Crosas, M. (2022) 'A Large-scale Study on Research Code Quality and Execution', *Scientific Data*, vol. 9, p. 60.

Turnbull, M. (2015) *Australian Government Public Data Policy Statement*, Department of Prime Minister and Cabinet, Commonwealth of Australia.

United Nations (2011) *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*, United Nations, New York. https://unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf

Vera-Toscano, E. and Wilkins R. (2020) 'Does Poverty in Childhood Beget Poverty in Adulthood in Australia?', *Breaking Down Barriers Report #1*, Melbourne Institute: Applied Economic & Social Research, Melbourne.

Vera-Toscano, E. and Wilkins R. (2022) 'Dynamics of Income Poverty in Australia', *Breaking Down Barriers Report #4*, Melbourne Institute: Applied Economic & Social Research, Melbourne.

World Economic Forum (2020) *Data Free Flow with Trust (DFFT): Paths towards Free and Trusted Data Flows*, World Economic Forum, Geneva. https://www3.weforum.org/docs/WEF_Paths_Towards_Free_and_Trusted_Data%20_Flows_2020.pdf

Zajac, T., Tomaszewski, W., Perales, F., and Xiang, N. (2021) 'Diverging Labour-market Trajectories of Australian Graduates from Advantaged and Disadvantaged Social Backgrounds: A Longitudinal Analysis of Population-Wide Linked Administrative Data', Life Course Centre Working Paper Series, 2021–21, Institute for Social Science Research, The University of Queensland.