Australian Economic Review

DATA ARTICLE OPEN ACCESS

The Melbourne Institute Data Lab, a Secure Access Environment for Informing Future Social and Economic Policy

Chaminda Rajeev Samarage | A. Abigail Payne

Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, Melbourne, Victoria, Australia

Correspondence: Chaminda Rajeev Samarage (rajeev.samarage@unimelb.edu.au)

Received: 15 February 2025 | **Revised:** 12 June 2025 | **Accepted:** 16 June 2025

Keywords: data sharing | secure access environment | social and economic policy

ABSTRACT

Policy analysts and academics play a critical role in informing policy design, implementation and evaluation. They apply their understanding of current social and economic issues, test theoretical frameworks and present new ideas that are a part of the ecosystem for promoting and sustaining efficient and equitable delivery of government programs. Enabling these roles through access, curation and analysis of data from multiple sources is a critical component of a well-developed analytic framework. This is particularly imperative as it relates to sensitive and proprietary data. Making these data more widely available unlocks many public benefits, but only if the risks associated with sharing data are properly managed. We introduce the Melbourne Institute Data Lab (the MIDL), a secure access environment that supports customisation of information security controls to balance between privacy and security, and user experience to support data-driven research for informing policy. MIDL is based in a university setting, which is an important feature given that universities are long-standing institutions that are independent and trusted for their endeavour to undertake non-biased research.

1 | Introduction

Over a decade ago, leading economists raised the importance of the emerging watershed of using big data for economic analysis. Varian (2014) highlighted the importance of building expertise in machine learning and collaborations with computer scientists and statisticians to harness the insights from big data for testing economic theory and undertaking better predictive analysis. In the same year, Einav and Levin (2014) discussed how big data, especially that related to harnessing large-scale administrative data sets and proprietary private sector data, could greatly improve what we measure for tracking and describing economic activities. These bigger and more extensive data sets permit researchers to better understand the consequences of events and policies.

The evolution of big data comes from the information that is collected and the ability to store a vast amount of information. Data today enable an analyst to stack information on an individual, household, community or organization to better assess the complexity of the opportunities with which we are presented and the actions we take. But what are the mechanisms we use to harness the opportunities with large sensitive data sets? And how do we enable broader access and collaboration that builds on the historic importance of universities serving as a safe haven for objective and impactful research?

This paper presents the Melbourne Institute Data Lab (MIDL), a secure access environment that provides data services for

Manuscript for consideration to the Australian Economic Review

statistical research to inform Australian policy by any researcher or analyst based in an Australian university.

Around the world, there are more and more papers making use of enormous data sets to inform and shape economic and social policy. To a lesser extent, Australia has joined other countries with the use of bigger data. Significant advances have included the Australian Bureau of Statistics development of data sets that can be linked to Census data (now known as PLIDA, personal level integrated data asset) (Parker 2017). There have also been other initiatives, such as the creation of the ATO Longitudinal Information Files (Abhayaratna et al. 2022).

Technological advances in computing power, data management and storage, and cybersecurity measures have resulted in an exponential rise in the amount of data available for micro-economic analysis, as well as driving the way we understand aggregate economies (Veldkamp and Chung 2024). Reports suggest that the amount of data in the world will grow from 44 zettabytes (ZB) in 2020 to around 175 ZB by 2025 (International Data Corporation 2018).[1] This increasing collection of data and the use of artificial intelligence to support a varying range of business and social needs, the amount of data consumed is continuing to rise. With generative AI such as Large Language Models (LLMs), scaling to increase parameters has led to models being trained on enormous amounts of data, with researchers suggesting that the current stock of data will be fully used by 2028 (Villalobos et al. 2024). There are different viewpoints in the field of economics on how to define 'Big Data' (Taylor et al. 2014), and there are opportunities alongside computer science for the design of new statistical algorithms for manipulating such data. But as Taylor et al. (2014) point out, larger data sets are critical for challenging the way economists think and apply economic perspectives in today's changing digital landscape.

While technological advances in computing power, data management, and cybersecurity measures have exponentially increased our ability to utilize sensitive unit record data for economic and social policy analysis, Australia lags behind many leading countries. There remains many data sets held by government, industry, and service organizations that are not accessible. And despite the groundbreaking work undertaken by organisations such as the Office of the National Data Commission, there is much more to do in Australia to create the environment for better evidence-based research and policy development and evaluation.

Australia has embraced protocols such as the 'Five Safes Framework' (Desai et al. 2016), which represents a significant advance for embracing increased access to sensitive data. Moreover, given the many uses for big data, Australia has accepted that a one-size-fits-all approach to data development and access is an undesirable solution.

When author Payne was in graduate school, a 'big' data set was one that captured 1000s of individuals or organizations over time. With such a data set, statistical power was an underlying concern. Today, with lots (hundreds of thousands, millions and billions) of observations and many covariates, while statistical power may not be an issue, a better understanding of the sources of variation and utilizing models that test for causality and not just significant correlations becomes paramount.

Although the importance of causality and rigorous modelling is addressed by more than academic researchers, the academy continues to play a critical role in the space of data creation, data curation and analysis. The evolution of data use for economic analysis in Australia remains in a nascent stage. One of the many reasons for the state of our work is tied to an underinvestment of resources to support secure data development, access and analysis. But what are we as university researchers doing to embrace what should be a data revolution in Australia?

This paper describes our contributions to support greater use of sensitive data by academics and the creation of opportunities for collaboration with policy analysts, government and industry for undertaking the analyses needed to support economic growth, wellbeing and addressing current social and economic issued faced by Australians.

Through a careful and strategic approach we acted on an existing need for housing data securely in the 21st century and transformed what was simply a secure environment for in house projects to create a state-of-the-art data lab that promotes better data access that is respectful of individuals and organisations that are studied, that approaches data custodians in a manner that addresses appropriate governance arrangements for granting data access, and that encourages collaboration across accredited data service providers (secure environments) and across research teams.

This paper highlights the importance of understanding that not all data sources are suitable for analysis and that most data sets require careful curation and transformation to permit quality economic analysis. Moreover, curation/transformation must evolve into processes that encourage efficiency (e.g. use of artificial intelligence and machine learning) but also the sharing of resources that recognises the importance of the public good nature of data. Essentially collaboration and sharing are vital if we are to identify Australia as an exemplar country for harnessing data for better economic and policy analysis.

The Melbourne Institute Data Lab's specific data services are catered to improve data users' experience with *access* to a range of data assets. The MIDL is suited for both standalone projects (single researcher or team) and for the development of themed *collaborative data environments* to permit wide-ranging access to connected data that have been curated and tests and the associated computer code and insights to enable faster and better analyses. Through collaborative data environments, we are tackling the challenges presented for appropriate governance structures that respect the individuals and organisations whose information is captured in the data sets, permitting study, but also enable a data custodian to retain control over who can be granted permission to access the curated data sets. A collaborative data environment, moreover, provides a necessary framework for enabling replication and testing of data quality, necessary elements for providing evidence-based insights to inform and shape policy.

Section 2 describes the recent developments for data development and sharing in Australia. Section 3 describes the Melbourne Institute Data Lab and the models it uses to support information security, privacy and data governance. Section 4 briefly concludes.

2 | Australian Context for Using Administrative Data

Gaining access to administrative data for analysis is often challenging and usually involves many layers of bureaucracy. While we have initially focused on data collected by the government (federal, state and local), administrative include data collected by industry and service organisations. In some parts of the world such as Nordic countries, government-based administrative data have been available to researchers for many years (Connelly et al. 2016). In Australia, access to and use of administrative data set remains underutilised but also fragmented due to complexities and variations in data sharing principles and laws across the Commonwealth, states and territories (Scheibner et al. 2023).

Australia's recent data strategies (2021 and 2023) have shown a commitment to supporting more secure data sharing and access to enable more data-driven research outcomes and to facilitate better policy outcomes for all Australians. Secure access environments (SAEs), or trusted research environments (Kavianpour et al. 2022), support access to sensitive micro-level data while upholding data confidentiality. The Australian Government's Data Sharing Principles, based on the Five Safes risk framework, an internationally recognised approach to managing privacy and disclosure risks, is a critical tool that data custodians and providers within government utilise to apply data protection controls surrounding research access to sensitive data assets.

Population data linkage is not a recent development. Back in 2009, the Population Health Research Network (PHRN) was established at the University of Western Australia to build and link data for the betterment of health and wellbeing in Australia. For more than a decade, it has been working with Commonwealth and state governments and other organisations. It is one of many pockets of opportunity that have been developed to encourage the building of information needed for good research and policy analysis.

As has been mentioned above, the Australian Bureau of Statistics and other Commonwealth departments, including the Department of Social Services and the Australian Tax Office, were also early investors in the creation of longitudinal and population-based data. At a state government level, most states have built linked data sets that are being used to assess state and regional service delivery.[2]

And yet, with all these developments, economic research that utilize these massive data sets is only now emerging. Recent legislative changes include the Commonwealth's Data Availability and Transparency Act 2022 (DAT Act), and these changes are catapulting the ability of researchers and analysts to easier access to Australia's high-quality public data sets, at least at a Commonwealth level.

These changes are enabling a level playing field for all researchers and analysts to gain data access. For example, the Data Availability and Transparency (DAT) scheme provides an accreditation process that enables a transparent and consistent approach for granting data access to researchers as well as an accreditation process for assessing a secure data environment for housing data (accredited data service provider) before they can access data from Government data custodians through accredited data service providers. The DAT scheme is implemented by the Office of the National Data Commissioner to ensure that the scheme is fairly and transparently implemented. Although the DAT scheme currently applies to academic researchers only. Access by private and not-for-profit sector analysts remains more limited.[3]

Another example of the potential to catapult empirical research using Australian data is the decadal vision for social science infrastructure report issued by the Academy of the Social Sciences in Australia (2024). This plan calls for action to setup an ecosystem for coordinating existing and emerging capability, ensure social science data are FAIR (Findable, Accessible, Interoperable, Reusable), and improve investment into the infrastructure sector (Wilkinson et al. 2016).

An emerging presence online, the use of digital services, and an ongoing need for data to train large Artificial Intelligence models such as Generative Large Language Models, is amplifying the need for secure environments to access sensitive data to inform future policy, while greatly minimising privacy risks to data custodians and data owners.

Underpinning most secure environments in Australia are a range of frameworks for protecting the confidentiality and privacy of hosted data. The Five Safes Framework provides a risk-based framework to minimise the risk of disclosure and re-identification of an individual, business or organisation. These principles are applied across five domains: projects, people, data, settings (the infrastructure used for data access) and outputs (any information taken out of the secure environment). This framework is also the basis for the Australian Data Sharing Principles that are enshrined in the Commonwealth's DAT Act. Other frameworks often used are the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al. 2016) and CARE (Collective benefit, Authority to control, Responsibility, and Ethics) principles proposed by Carroll et al. (2021) for Indigenous data governance relate to the people and the purpose of data.

These frameworks were not intended to be stand-alone but were designed to complement proper data governance and stewardship. From a security perspective, there are different recognised industry information security and technology standards that are applicable. This includes state-level frameworks such as eHealth NSW (Privacy Security Assurance Framework [PAF]), Commonwealth policies (Protective Security Policy Framework [PSPF]), and Information Security Manual (ISM) assessed under the Information security Registered Assessor Program (IRAP) assessment, the Essential 8 Maturity Model to international standards and frameworks (ISO/IEC 27001:2022 and NIST SP 800-53).

Our work through the MIDL highlights that there is no 'one-size-fits-all' solution. There are differences in the requirements of data users based on their level of data analytics expertise and their reasons for the level of analysis they want to undertake. High-end data users typically want to extract full value from microdata (i.e., data with a unit of observation at an individual, household, or organisation level). This is often through a real-time, interactive environment, which in the literature is often referred to as secure access environments (SAEs), trusted research environments (TREs), digital research environments (DREs), virtual data enclaves or virtual research data centres (Kavianpour et al. 2022). Irrespective of nomenclature, these environments allow approved researchers to

access sensitive micro-level data with sufficient information security controls to protect the privacy of the underlying data units.

The MIDL is adding to the existing social science research infrastructure in Australia. The MIDL's environment provides researchers access to statistical analytical software packages such as Stata, R, Python and systems that permit them to use application program interfaces (APIs) to bring other data into the environment. An interview study of platform providers (Kavianpour et al. 2022) suggested that further work for next generation environments should provide features such as the ability to provide analytical power for large-scale studies; bringing data, algorithms and code into the environment safely; ability to develop artificial intelligence and machine learning models within the environment and export; supporting queries of data from external sources; simplifying the process to access data by researchers; and scalability as the number of data users grow.

The MIDL's set-up has resulted in it becoming Australia's first university-based Accredited Data Service Provider (ADSP) under the Commonwealth DAT Act. Yet, the MIDL is emerging as more than an ADSP. Its aspirations are to be more than what might be historically referred to as a 'data warehouse'. Through the work of researchers and data analysts that use the MIDL and the creation of collaborative data environments, the MIDL is a source for testing and curating data before analysis that will result in researchers being able to access the curated data through the MIDL. Effectively, the MIDL is designed to support better, faster, and deeper research analysis.

The MIDL is not an island. For Australia to become a leader in the use of national and proprietary data sets for evidence-based innovation and policy analysis, there is an importance of promoting collaboration across other secure data enclaves across universities, governments, and industry. We have built the MIDL and are engaging in projects that will showcase the importance of interoperability between cloud environments (Grossman et al. 2024). Future work will explore how different like-for-like cloud environments can collaborate with each other – improving the overall user experience when working across different research environments. This study also aims to navigate the potential challenges presented by complex regulations between state and federal legislation, such as that observed in Australia. This study will be further extended with collaborators at other universities and will include addressing the use of private and service provider data in secure environments designed for research and innovation.

3 | The Melbourne Institute Data Lab

The Melbourne Institute Data Lab (the MIDL) provides a secure environment that facilitates the development of collaborative and coordinated data environments for authorised researchers and analysts, allowing the rigorous and deep study of critical issues important to Australian society through data access, curation, and the sharing of knowledge about the veracity of the data. The MIDL is managed by the Melbourne Institute of Applied Economic and Social Research, an independent research department within the Faculty of Business and Economics at the University of Melbourne. The MIDL serves as an Accredited Data Service Provider for all University of Melbourne researchers and has the capacity to enable use by researchers at other universities and permits a platform for stronger engagement and collaboration with analysts and decision makers. The MIDL is not just for writing academic papers. The MIDL is an environment that permits enabling deeper engagement with data (e.g. through the offering of curated data and the development of interactable visualisations on specific topics) and provides a platform for testing ideas for encouraging innovation in the data curation and data analysis space.

The MIDL platform is the result of an industry collaboration between the Melbourne Institute and Cyconsol, an Australian-based professional services provider specialising in the design and operation of cybersecurity and ICT capabilities working with the Australian Government and industry.

The MIDL has been designed by researchers, specifically for researchers and data custodians. It builds on the previous secure environment developed by the Melbourne Institute (the Melbourne Institute Secure Environment) that was used for housing the production of the Household, Income and Labour Dynamics (HILDA) Survey for the Department of Social Services and other data products throughout the longstanding history of the Melbourne Institute.

The MIDL platform:

- Supports the Australian Data Strategy (Department of Prime Minister and Cabinet 2021) and the Australian Data and Digital Government Strategy (Department of Finance 2023).
- Supports data sharing under the relevant laws and legal instruments including the Data Availability and Transparency Act 2022 (DAT Act).
- Aligns with the Five Safes Framework and the Australian Data Sharing Principles.
- Is an Accredited Data Service Provider under the Office of the National Data Commissioner's Data Availability and Transparency (DAT) Scheme.
- Continues to provide assurances under the Australian Government's Protective Security Policy Framework (PSPF) and the Australian Cyber Security Centre's Information Security Manual (ISM) through regular assessments under the Australian Signals Directorate's Infosec Registered Assessors Program (IRAP).

In line with the mission of the Melbourne Institute, the primary aim of the MIDL is to enable high-quality academic or industry engagement research on major economic and social policy issues affecting contemporary Australia. The objectives of the MIDL platform are to:

- Provide a collaborative data environment for use by authorised researchers and analysts, permitting rigorous and deep study of critical issues important to Australian society.
- Enable access of administrative data sets from Government and proprietary sources, and data collected through surveys and field experiments, to trusted users.

- Enable remote access through a secure virtual desktop environment and provide analytical capability for users to conduct their analyses.
- Enable data custodians to provide data with maximum utility for users while maintaining the confidentiality of the information.
- Provide a platform to enable data curation through the ability to share additional material from the MIDL staff, the data custodian, and researchers.

In accordance with protective security requirements outlined in the Australian Government's Protective Security Policy Framework (PSPF) and Information Security Manual (ISM), the MIDL was built to meet the requirements for housing PROTECTED information. The MIDL was designed with a higher security posture to enable data users with the leverage needed to assure custodians of sensitive data, enabling greater access and increased utility for research without compromising data privacy and confidentiality.

3.1 | Technical Implementation

The MIDL leverages a virtual desktop infrastructure to enable approved data users with remote access to a virtual desktop environment, enabling access beyond a locked room in a single building. This environment provides appropriate software and the capability to develop, test and execute analyses of these data to answer their research questions. The MIDL provides each data user (or sometimes shared across a group of data users) with their own Virtual Machine (VM), or virtual desktop (see Appendix), with fixed resources (memory, processing power and storage). These virtual desktops are accessible over an encrypted internet connection and provide users with a range of statistical analytical packages and programming languages for data analysis and visualisation. The MIDL's virtual desktops can be accessed remotely using a web browser (see Figure 1), and authentication hardening is achieved through the requirement of multi-factor authentication on both user and system administrator (accounts which have elevated rights to make changes to the environment) accounts used by personnel who maintain the secure infrastructure.

Data provided by data custodians (research data) and data and other artifacts created through analyses by approved data users (project data) cannot be removed from the MIDL's systems. Output of analyses is vetted before release from the MIDL environment.

Data users are only allowed to view and analyse approved data assets using a virtual desktop. All data held within the MIDL's system is encrypted using AES-256 encryption. Analytical outputs must be vetted before leaving the MIDL environment. The vetting

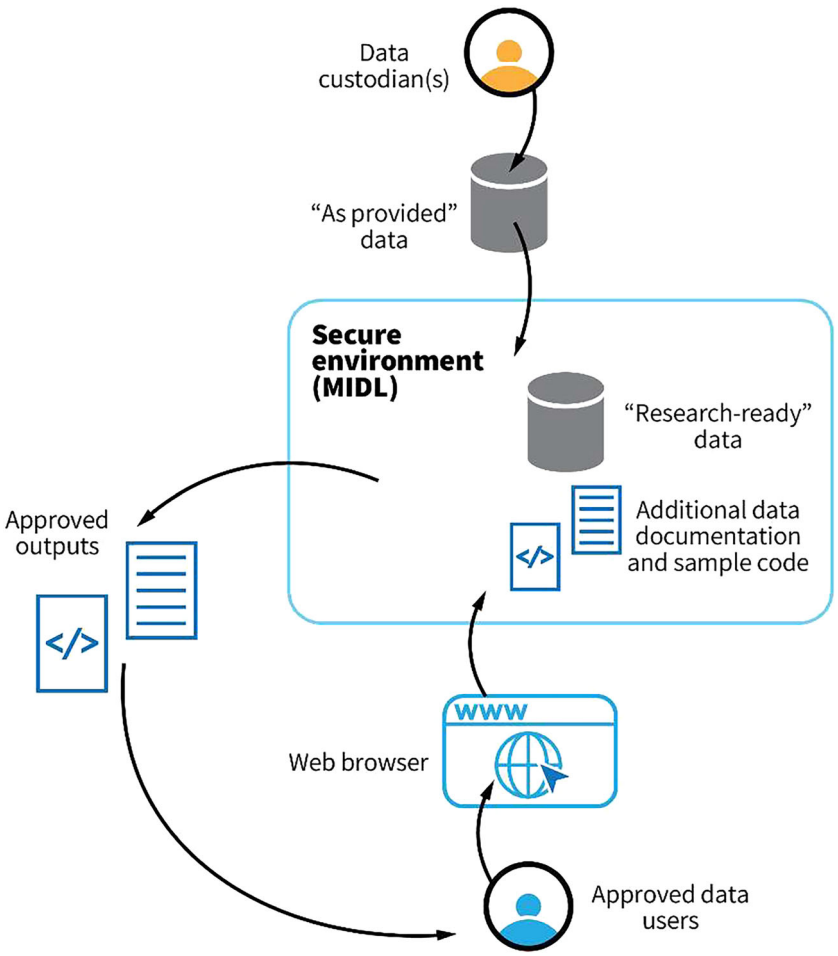

FIGURE 1 | Inputs and outputs from the MIDL environment.

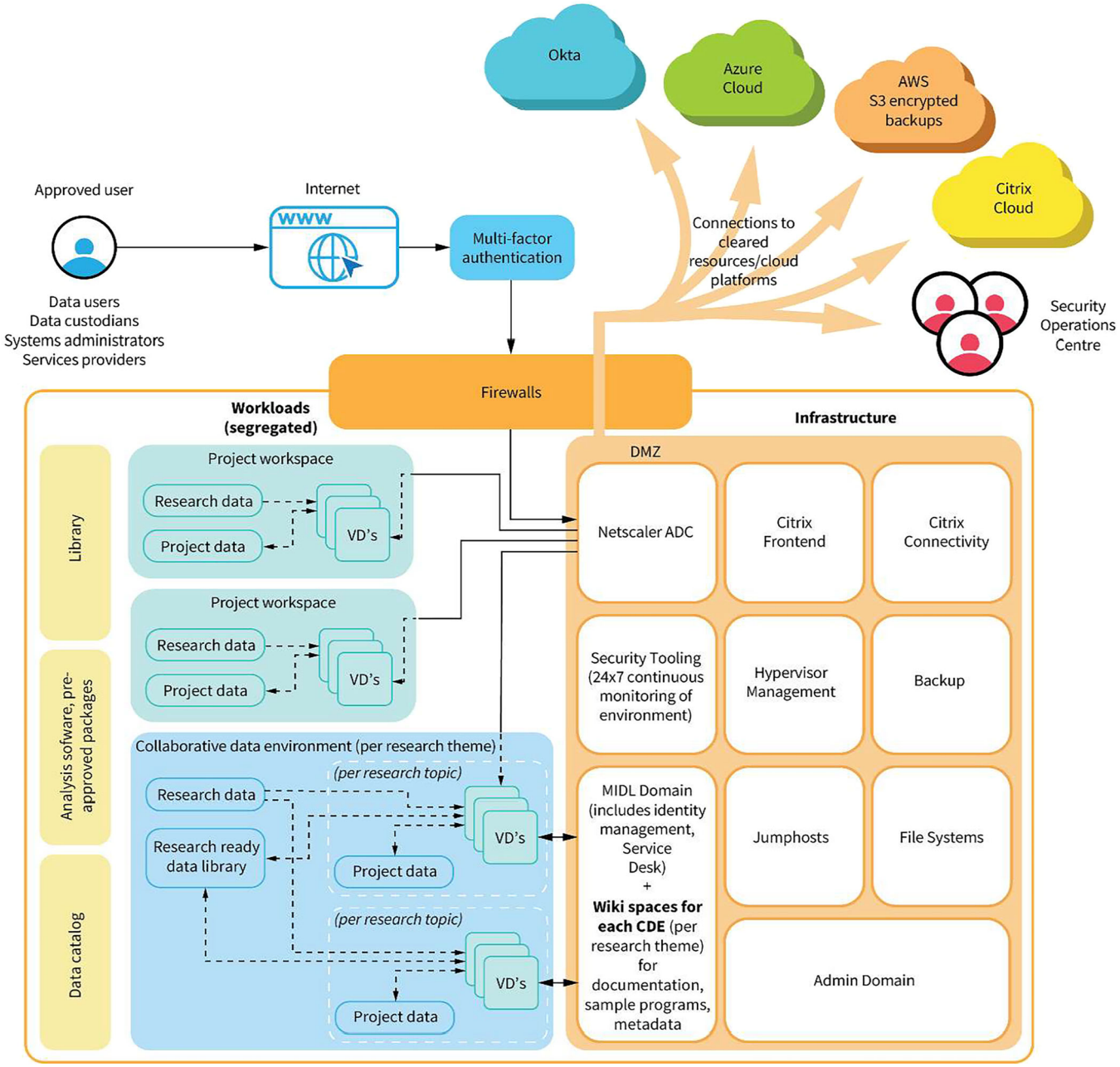

FIGURE 2 | The MIDL architecture diagram. Abbreviations: ADC, application delivery controller (product name for part of virtual desktop delivery solution); AWS, amazon web services cloud infrastructure; CDE, collaborative data environment (see Section 3.3 for more information); DMZ, de-militarized zone (a network segment that acts as a buffer between the private infrastructure and an untrusted network i.e. the internet); VD, virtual desktop (see Appendix).

protocols are set by each data custodian. All protocols follow a manual, principles-based vetting process that is managed by the Melbourne Institute's Data & Analytics Team. These principles are based on best statistical disclosure control practices to evaluate disclosure risks and apply appropriate confidentiality measures as required before being released outside the MIDL environment.

Necessary steps have been taken to minimise the risk of unauthorised data disclosure outside of the MIDL environment. For example, the copy/paste functionality is disabled, preventing a user from copying into or out of a virtual desktop. Moreover, there is no internet, e-mail or team chat capability inside a virtual desktop. Subject to data custodian approvals, specific internet addresses can be whitelisted at a project level to enable access to online data dictionaries or application program interfaces (API's) to bring in approved data for further analysis work. Effectively, the MIDL provides a platform where risk-accepted approaches can be applied to permit the provision of custom controls at a project level, depending on data custodian approvals and user requirements.

The MIDL environment is hosted within a secure data centre based in Victoria, Australia and primarily leverages (see Figure 2 for MIDL architecture diagram) Nutanix hypervisor technology and Citrix virtual desktop architectures to provide data users with the ability to securely access environments to

undertake data analyses on sensitive data assets provided by data custodians. The design of the MIDL is separated into two main components:

- **Workloads** that house projects and collaborative data environments that enable research activity by data users of the MIDL; and
- **Infrastructure** that makes up shared resources for the MIDL that provide administrative, operational and security capability for the MIDL environment.

A key aspect of the infrastructure that supports the MIDL's collaborative data environments is the MIDL Wiki. The MIDL Wiki is a dedicated on-premises instance of Atlassian Confluence with customisations to set up entire wiki spaces for different data assets and collaborative data environment home spaces. Permissions are defined using role-based access control groups that limit data users' views to allowed data spaces based on the project that they are using to access the wiki. For example (these do not reflect actual projects) assume Project A contains Australian Census data and assume Project B does not contain these data. Project A users will be able to utilise the wiki space for the Australian Census, but Project B users will be denied access to this wiki space.

The MIDL's infrastructure also relies on several cloud-based solutions. This includes Okta Cloud for multi-factor authentication and identity management. The MIDL undergoes regular vulnerability scanning and 24/7 security monitoring through the MIDL's Secure Incident Event Monitoring (SIEM) system with security operations capability provided by Cyconsol and a cleared third party based in Canberra, Australia. This SIEM system continuously scans the MIDL's network infrastructure, monitors network traffic including logins and data sent into and out of the environment, execution of programs and scripts from within the environment, for usual and unusual behaviour patterns. Backups of research data, project data and system data are carried out daily to an on-premises solution and backed up weekly to the MIDL's own cloud instance on Amazon Web Services. By default, the MIDL has a data retention policy of 5 years with the ability for project leads to specify different data retention periods as required by the data sharing agreements in place.

3.2 | Alignment With the Five Safes Framework

The MIDL's onboarding, access, and output processes have been designed to align with the Five Safes Framework: Safe People, Safe Projects, Safe Data, Safe Settings and Safe Outputs. The authors and the MIDL team opted to use this model as it is closely aligned with how data sharing is undertaken, and more recently written into law through the DAT Act as the Australian Data Sharing Principles. This model allows the MIDL to minimise and manage disclosure risks and enable trust among data custodians that data held within the MIDL and used by approved data users are managed appropriately. This approach enables the MIDL team (and where applicable, data custodians) to appropriately assess projects, users, data provided for research purposes, and the outputs that are created as part of research activity.

Table 1 provides an overview of how the MIDL effectively manages disclosure risks through the application of the Five Safes.

3.3 | Access Methods

The MIDL allows users to access data in two ways.

- Through the setup of a **standalone project**, or
- Through access to a **collaborative or coordinated data environment**.

Standalone projects allow approved data users to undertake research activity on approved data sets to answer research questions as stated in their project applications that have undergone vetting. Standalone projects also incorporate the housing and analyses of primary data collection (e.g. surveys, field experiments). Within the MIDL policy framework, standalone projects, thus, utilise a 'bring your own data' (BYOD) model. In this instance, project leads are provided with limited administrative access to folder permissions that allow them to set the rules of use of these data.

Collaborative and coordinated data environments represent a collection of projects that fit under a broader research theme. A collaborative environment would reflect researchers or teams of researchers actively engaging with each other to create an environment to ensure that relevant data sets are brought in and curated in a manner to enable the effective use of each data set, to address issues that arise that relate to the collection of information and changes to the collection of information that make up the data set, and the ability to link measures across data sets. A coordinated data environment is one that also involves multiple researchers and research teams but where the teams operate more independently but in a coordinated fashion to support each other's work.

The environments enable researchers to undertake a series of projects that investigate a range of research questions that fall under the theme of the collaboration (e.g. disadvantage). Collaborative and coordinated approaches encourage efficiencies in the undertaking of data curation that results in greater consistency in the underlying data.

A collaborative or coordinated environment, however, does not mean there is automatic universal access to all data housed in the environment. Following the governance standards that reflect the importance of the data custodian's role for enabling research analysis, access to the data sets for any given project must still be negotiated with the relevant data custodians. This allows the data custodian to define its own conditions for use of the data within the collaborative environment. This governance framework ensures that each project utilising the environment is given access only to the set of data sets that have been approved by the relevant custodians. The use of the MIDL and a collaborative or coordinated approach, however, supports data custodians in terms of their being able to release their data to a single location and to develop consistent practices for granting permission to use the data across projects.

The value of the collaborative or coordinated data environment is that researchers focused on a topic for the collaboration can

TABLE 1 | Application of the five safe risk framework in the MIDL.

Safe	How it is applied to the MIDL
Safe people	Users must undergo security awareness training and complete an authorisation process before data access. A user's organisation(s) must first sign Access Agreements (head agreement) between the organisation and the University of Melbourne for the MIDL access. At a project level additional project schedules including the MIDL Acceptable Use Policy, and any additional terms set out by the data custodian, must be signed before being granted the MIDL access. This project schedule includes a confidentiality agreement that stipulates a user's requirements to maintain data privacy and confidentiality whilst using the MIDL. Users must also undertake an annual MIDL security awareness training module (online) which covers shared responsibilities of user's in maintaining data security and data confidentiality of assets held within the MIDL. The MIDL training also covers basic concepts in statistical disclosure control to ensure data privacy is maintained when taking data outputs out of the secure environment. Data custodians may elect to provide additional data-specific material that forms sub-modules that must be completed by users wanting to access a particular data set. This training also includes key contact details for raising service requests and security incidents. Data custodians may elect to nominate a user approval process that requires additional security controls such as police checks and/or security clearances before granting access to their data sets hosted within the MIDL.
Safe projects	Current processes require project leads to: • Ensure their project is of public interest and for research or statistical purposes. • Describe their project and project objectives. • Provide details on partner organisations and/or funding. • Provide details on ethics approvals for project activities. • Provide details on required data assets. • And provide additional security controls (include data retention policies) that may be applicable to their project through additional requirements from any relevant data sharing agreements. When establishing the data sharing agreements, processes can be put in place to allow data custodians to nominate an approval process for users and/or projects using the information captured above. This would allow data custodians to enquire on how users' intend to use the data and share outputs. This requirement ensures that the project aims align with a data custodian's mission statement and values.
Safe settings	Information security controls implemented within the MIDL for a security classification of PROTECTED under Australian Government Protective Security Policy Framework and the Information Security Manual (ISM). The MIDL utilises an array of information security controls as required by the Australian Government and international regulations. This includes: • Strong authentication and identity management controls, including the enforcement of multi-factor authentication protocols. • 24×7 ongoing security monitoring of facilities and environment (SIEM/SOC). • Citrix remote desktop environment with additional security controls to ensure data confidentiality is maintained. • Data storage and archiving with retention up to 7 years (extendable to 20 years). • Citrix-based solution for file ingress/egress vetting and approvals before files/outputs can be transferred into or out of the MIDL environment. Under the MIDL's Assurance and Audit Plan, the MIDL undergoes the Australia Government's Information Security Registered Assessor's Program (IRAP) by an external auditor every two years. The MIDL also undergoes annual penetration testing by an independent services provider. The MIDL is an Accredited Data Service Provider (ADSP) under the Office of the National Data Commissioner's Data Availability and Transparency scheme to provide data services and enable data sharing for data assets held by the Australian Government.
Safe data	The MIDL accepts data that have direct identifiers removed with further treatments applied to minimise disclosure risk. Where this is not applicable, the MIDL will leverage its increased security posture with additional security checks for users who need access to sensitive information. Data containing direct identifiers will not be released to the user from inside the secure environment.

(Continues)

TABLE 1 | (Continued)

Safe	How it is applied to the MIDL
	A range of statistical disclosure control tactics are highlighted in the MIDL security awareness training as techniques that users can implement on their outputs before preparing for vetting by the data custodian. For more information, please see Data Management and Governance Processes.
Safe outputs	All statistical outputs and visualisations will need to be vetted and cleared before they are taken outside the MIDL environment. This is implemented using a Citrix-based custom solution that allows data custodians to view outputs prepared by the users and approve/reject or provide feedback to ensure statistical results are non-disclosive. Material in the MIDL Wiki cannot be taken outside the MIDL environment without approval from the data custodian(s). There is no copy/paste functionality, print capability or the use of removable media storage or internet functionality inside the MIDL environment to ensure data cannot be taken out without the use of the MIDL output vetting processes.

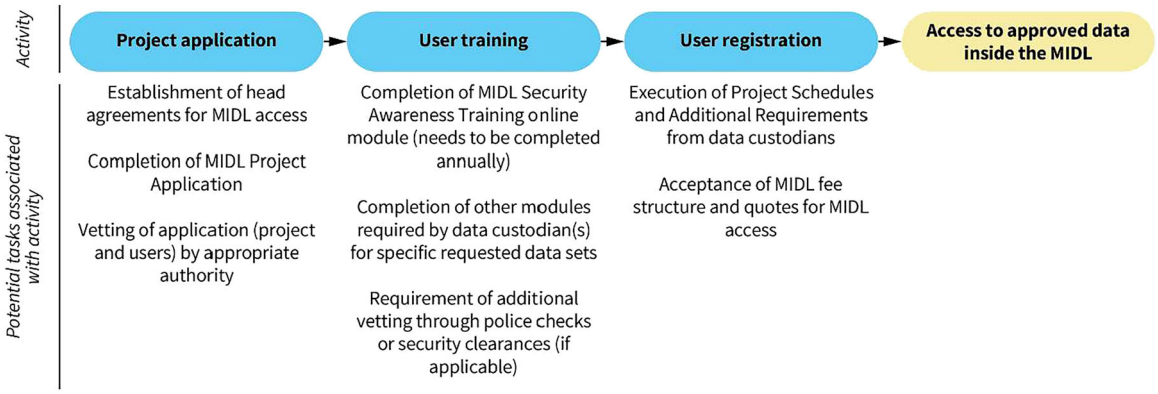

FIGURE 3 | The MIDL user onboarding journey.

share important information and curation techniques with other researchers. This sharing and joint curation approach enables greater consistency in data quality as well as promotes more timely analyses.

We see collaborative and coordinated data environments evolving to serve as 'digital communities of practice' where data curators work with the data custodians to create 'research ready' versions of data sets research and policy analysis.

Long-term, research-ready data represent a unique opportunity to reduce duplication of effort in curating data for research. The authors define research-ready data as FAIR, i.e. that they are Findable, Accessible, Interoperable and Reusable for research purposes. Key characteristics of research-ready data can be found in Mc Grath-Lone et al. (2022). A recent study (Perry and Netscher 2022) found that up to 5000 h can be spent on cleaning the data and updating data documentation for data re-use, noting that this study used curators who are experts in cleaning and documenting data. It is very likely that individual researchers from non-technical backgrounds may spend more time on data curation. To allow more customisation options, collaborative data environments can be configured with its own governance structures for project approvals, user vetting, and developing data acquisition plans by its project leads.

Irrespective of which way data are accessed, all potential data users must go through the MIDL user onboarding processes (blue tiles depicted in Figure 3). This process typically takes 5 working days (subject to quick completion of training modules), and approved data users are contacted via email. The MIDL's onboarding process consists of the following processes.

- **Project application:** Prospective data users must complete a project application to access the MIDL environment and access data through a standalone project or a project within an existing Collaborative Data Environment. This project application is either vetted by the MIDL services team or delegated authorities of the data custodian(s). A range of criteria such as the project's purpose, requested data assets, intended outputs and ethics approvals play a key role in project vetting. Other characteristics such as data user capability (data analytics expertise, prior history of data access and working with data, other data training such as on other data environments) is also considered in cases where this information is available. By default, projects intending to utilise the MIDL must have a purpose or intention to undertake statistical research for informing Australian social and economic policy.
- **User training:** All prospective data users must complete an online-based security awareness training. Users must continue

to complete this module every year, as this is a regulatory compliance requirement for the MIDL. This module focuses on shared user responsibilities (both legal and ethical) and expectations to minimise the disclosure of data privacy while using the MIDL service. Data custodians have the option to provide additional training material that users need to complete before accessing their data assets. Services providers and system administrators for the MIDL infrastructure must complete additional user training in addition to the standard training before being granted access to the MIDL's administrative functions. Depending on the level of access, all systems support staff must provide recent Australian Police Checks, and administration personnel must hold an Australian Government Security Clearance of Baseline for ICT operations or Negative Vetting 1 for security operations.

- **User registration:** Once training has been successfully completed, all approved data users must complete and sign project schedules and individual undertaking to access requested data inside the MIDL. Data custodians could dictate additional requirements for data users requesting their data. This enables data custodians with the flexibility to 'pass through' legislative requirements for data access to the data users.

3.4 | Using the MIDL

Approved data users can remotely access the MIDL environment from anywhere in Australia. Additional steps such as data custodian approval can be sought for the MIDL access from pre-approved locations across the globe. Data users use their credentials (multi-factor authenticated username and password) to access a virtual desktop using a web browser on their workplace or personal computer. The MIDL User Guide provides data users information on how to prepare their personal computers to access the MIDL. Before access all data users and informed that they must agree to the MIDL's acceptable use policy.

Once successfully authenticated, data users can select which approved project they can access inside the MIDL. Clicking on a project icon takes the user to a virtual desktop screen (see Appendix) that is enabled by the MIDL's local Citrix-powered virtual desktop infrastructure. This virtual desktop enables user to access their approved data assets in an environment that provides licenses statistical analytics software and programming languages.

3.5 | Storage Drives

Data users have access to four storage drive mappings inside the MIDL for different purposes. The 'D: DATA' drive provides access to approved data sets available for the project. In projects where data is brought into the MIDL under the BYOD model, project leads have limited read/write capability to the D: drive to set up folder structures and folder permissions for other listed data users on their project. In all other instances, the data folders within the D: drive are read-only. The D: drive also hosts the 'DATA CURATION' folders used for file transfer into and out of the MIDL environment.

The 'H: HOME' drive contains the data user's home folder that is specific to their MIDL account and the project they are accessing. This means that data users are not able to copy sensitive files between projects and other data users cannot see the contents of these folders. This folder is regularly backed up and archived and can be used for project activity. The 'W: PROJECT WORKING' drive is the recommended folder for sharing data analyses and documents resulting from project activities with other approved data users of the project. All data users in the project will have read/write access to this folder, unless specifically controlled by the project lead for other administrative reasons.

The 'L: LIBRARY' drive is a read-only resource folder that is viewable by all data users inside the MIDL. It provides access to some openly available resources, such as Australian geographical boundaries from the Australian Bureau of Statistics (ABS), and any other relevant analytical reference material, as there is very limited access to the internet from inside the MIDL. The MIDL allows data custodians to dictate limited internet access through their data sharing agreement with the MIDL. For example data custodians may opt to place specific parts of a website (such as the ABS Census Dictionary) in a Uniform Resource Locator (URL) allow list that is configured at a user or project level in the MIDL's firewall configurations. This would allow approved data users access to limited parts of the internet from specific pre-approved projects, which limits the need for a broader library resource folder.

3.6 | Analytical Software and Programming Languages

As the MIDL services data users with broad and extensive expertise in data analytics, the MIDL provides a range of software packages and programming languages for statistical analysis. These include the latest versions of Stata/MP, SPSS, SAS, StatTransfer, and programming languages such as R using the RStudio integrated development environment and Python through the open-source data science platform, Anaconda. The MIDL also provides data users with popular packages from the Microsoft Office suite of products and PDF viewing capability through Adobe's Reader package. Existing users of Adobe Creative Cloud can request access to Adobe packages but access to cloud resources such as file sharing have been switched off to prevent unauthorised egress of data from the environment.

Packages signed by StataCorp can be installed using the 'ssc install' command. These packages are installed directly to the data user's home drive and will be available on the same project the next time they login. Additional packages for Stata, R and Python can be added by directly contacting the MIDL services team. All software package requests are vetted by the MIDL team and any associated software costs will be passed as an increase to their data access fees.

3.7 | Support and Service Desk

The MIDL's service desk is operated by dedicated Melbourne Institute staff, dedicated Cyconsol operations, engineering and security personnel, and members of the Melbourne Institute Data & Analytics team. The Melbourne Institute operate a dedicated email (MIDL-info@unimelb.edu.au) that lets

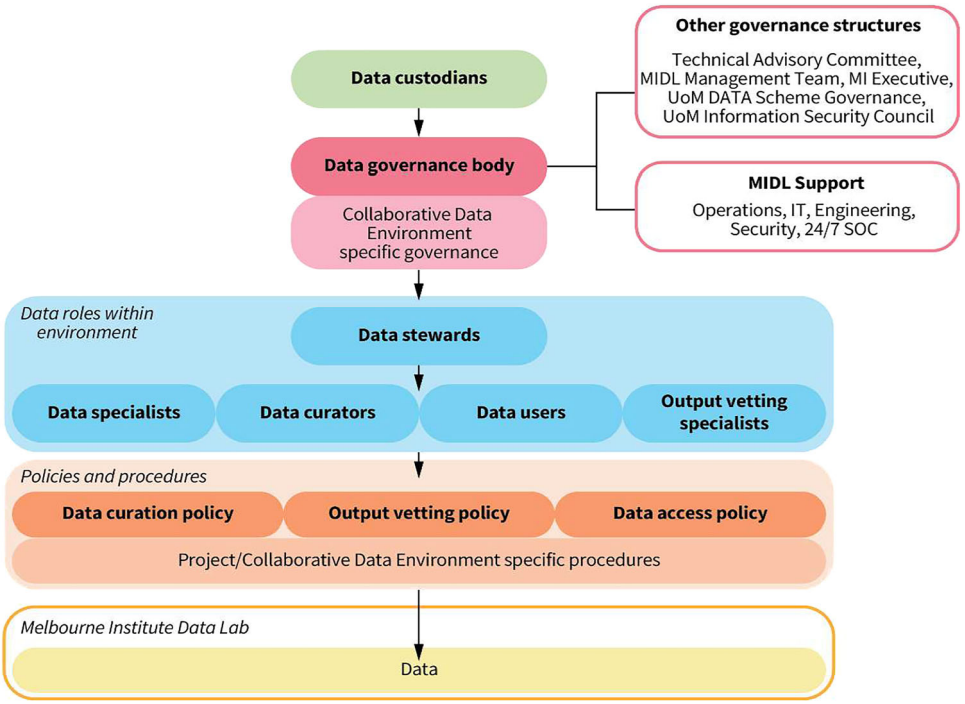

FIGURE 4 | The MIDL data management model.

potential data users to contact and get further information about the MIDL and its services. The MIDL team also operate a secondary email (servicedesk@MIDL.unimelb.edu.au) for directly contacting MIDL's service team. This email is limited for existing and previously approved MIDL users and is controlled by MIDL's firewalls. This email is configured to an IT service management (ITSM) system that creates a service ticket for each email. The MIDL's support is tiered to meet the needs of the ticket with support ranging from operational support, IT support, IT engineering and security and data analytics. Data analytics support includes output vetting requests, queries related to the use of software packages, and vetting of new software packages to be installed.

3.8 | Output Clearance Processes

Output vetting and clearance are key requirements under Five Safes Framework. It is a mechanism to minimise statistical disclosure and re-identification of units (i.e. individuals, households, businesses) within a sensitive data set. The key service offering of an SAE is access to data that has undergone minimal treatment to prevent data disclosure, such as data aggregation. Output vetting enables data custodians to manage disclosure risks and any mistakes by data users with limited expertise in working with sensitive data.

The MIDL's input and output vetting processes are implemented using custom workflows within Citrix's own content collaboration tool, Citrix ShareFile. Data users wishing to take project outputs outside of the MIDL must place their files in the 'D: DATA drive' under 'DATA CURATION/OUTPUT' folder. This triggers an automated message to the project's nominated output vetting specialist(s). Data users are also encouraged to raise a service ticket by emailing all relevant details of the output vetting request directly to the MIDL's service team. The MIDL's custom workflow and service ticketing system through the MIDL ITSM lets data users engage in a conversation with the output vetting specialist and address any concerns raised.

Output vetting of data inside the MIDL may be conducted by MIDL's own output vetting specialists as delegated by data custodian(s) or staff who are employed by the data custodian(s). These details are established as part of the data sharing agreement between the MIDL and the data custodian(s). Where applicable, a 'principles'-based approach is used to vet output, and these are broadly defined in the MIDL Output Vetting Policy. This policy establishes output vetting rules based on the sensitivity of data. General rules under this policy include base rules (to check for identifiable information in the output, bottom coding, and top coding of extreme values) and other rules such as those used by the Australian Bureau of Statistics' ABS DataLab (rule of N, dominance rules, group disclosure rules). Specific rules under this policy dictate additional output vetting rules by specific output types and are derived from Bond et al. (2016).

Once the outputs have been cleared for release, they are released for the data user to access from the MIDL's Citrix ShareFile gateway outside of the secure environment (data user credentials are needed for access).

3.9 | Data Management and Governance Processes

Within the MIDL platform, the ICT procedures, administrative operations, and security operations including ongoing regulatory compliance activities are handled through the MIDL Governance Framework. The scope of this framework is for activities around the three following domains:

1. Regulatory compliance;
2. Data; and
3. Governance templates for Collaborative Data Environments hosted within the MIDL.

Through the MIDL Governance Framework, there are several governance structures to ensure appropriate oversight, roles and responsibilities and approval mechanisms are in place to achieve the above. These include committees and teams for change management, incident response, administration and day-to-day operations, project and user application vetting including vetting of administrator users; audits of projects, users and data assets, to name a few.

Data management within the MIDL is overseen by the MIDL Data Governance Body and data governance activities are defined in the MIDL Data Management Model (see Figure 4). This model ensures that access to data, curation of data and vetting of data to be taken outside the secure environment are managed behind layers of security, protocols and procedures, and policies that are governed by different roles.

4 | Conclusion

The Melbourne Institute Data Lab is a university based secure access environment that has been developed with a strong focus on providing an environment that meets needs of researchers and data custodians efficiently and securely to enable the development of research, the creation of innovative statistical and data techniques, the forming of evidence-base analysis to shape and inform Australian economic and social policy. It also defines two frameworks for data access through standalone projects and through a collaborative or coordinated data environment, a digital community of practice that supports projects to be set up under a broader research theme. The MIDL team, and its larger team of service providers, have helped build an environment that can adapt to the rapidly changing cybersecurity landscape and data sharing policy landscape in Australia while providing customisations to meet data researcher and data custodian needs.

Acknowledgements

Open access publishing facilitated by The University of Melbourne, as part of the Wiley – The University of Melbourne agreement via the Council of Australian University Librarians.

Data Availability Statement

For more information on the MIDL including its data services and the process of gaining access, please contact the MIDL team via email at MIDL-info@unimelb.edu.au. More information on the MIDL is also available on https://melbourneinstitute.unimelb.edu.au/data/MIDL.

Endnotes

[1]One zettabyte is equal to one billion terabytes.

[2]Several bespoke environments have been established to host sensitive data to support medical research (SeRP Australia, Secure Unified Research Environment 2023), non-health related disciplines (Parker 2017) as well as supporting hosting of a hybrid of health and non-health data assets (KeyPoint Secure Vault 2023; E-Research Institutional Cloud Architecture 2023).

[3]The Data Availability and Transparency Act (Cth, 2022) provides a framework through which legislative barriers for data sharing by data custodians within Australian Government are lowered. This is a significant step forward, but the accreditation scheme and regulatory complexities may discourage data sharing as it has been observed elsewhere across Australia (Scheibner et al. 2023). At the time of writing, this Act only allows Commonwealth, state and territory government bodies, and Australian universities to participate in this scheme limiting the extent to which independent research institutions and organisations can participate. A recent 5-year productivity inquiry report (Productivity Commission 2023) by the Productivity Commission recommended that the DAT Act be extended to include data sharing with accredited organisations in the private sector, such as businesses and not-for-profit organisations.

References

Abhayaratna, T., A. Carter, and S. Johnson. 2022, "The ATO Longitudinal Information Files (ALife): Individuals—A New Dataset for Public Policy Research." *Australian Economic Review* 55: 541–557. https://doi.org/10.1111/1467-8462.12486.

Academy of the Social Sciences in Australia. 2024. "Connected, Innovative and Responsive: Decadal Plan for Social Science Research Infrastructure 2024–33." https://doi.org/10.60651/90pr-cz87.

Bond, S., M. Brandt, and P.-P. de Wolff. 2016. "Guidelines for Output Checking." Technical Report European Commission, FP7-SP4 Capacities, Project number 262608, Data Without Boundaries.

Carroll, S. R., E. Herczog, M. Hudson, K. Russell, and S. Stall. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous Data Futures." *Scientific Data* 8, no. 1: 108. https://doi.org/10.1038/s41597-021-00892-0.

Connelly, R., C. J. Playford, V. Gayle, and C. Dibben. 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research* 59: 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Department of Finance. 2023. *Data and Digital Government Strategy: The Data and Digital Vision for a World-Class APS to 2030*. Department of Finance, Commonwealth of Australia.

Department of 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.

Desai, T., F. Ritchie, and R. Welpton. 2016. "Five Safes: Designing Data Access for Research." Economics Working Paper Series, 1601, 28, Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England, Bristol.

Einav, L., and J. Levin. 2014. "The Data Revolution and Economic Analysis." *Innovation Policy and the Economy* 14, no. 1: 1–24. https://doi.org/10.1086/674019.

E-Research Institutional Cloud Architecture. 2023. University of New South Wales. Accessed November 9, 2023. https://research.unsw.edu.au/erica.

Grath-Lone, L. M., M. A. Jay, R. Blackburn, et al. 2022. "What Makes Administrative Data 'Research-Ready'? A Systematic Review and Thematic Analysis of Published Literature." *International Journal of Population Data Science* 7, no. 1: 1718. https://doi.org/10.23889/ijpds.v6i1.1718.

Grossman, R. L., R. R. Boyles, B. N. Davis-Dusenbery, et al. 2024. "A Framework for the Interoperability of Cloud Platforms: Towards FAIR Data in SAFE Environments." *Scientific Data* 11: 241. https://doi.org/10.1038/s41597-024-03041-5.

International Data Corporation. 2018. "Data Age 2025: The Digitization of the World From Edge to Core." IDC White Paper #US44413318, International Data Corporation, Massachusetts.

Kavianpour, S., J. Sutherland, E. Mansouri-Benssassi, N. Coull, and E. Jefferson. 2022. "Next-Generation Capabilities in Trusted Research Environments: Interview Study." *Journal of Medical Internet Research* 24, no. 9: e33720. https://doi.org/10.2196/33720.

KeyPoint Secure Vault. "Queensland Cyber Infrastructure Foundation." Accessed November 9, 2023. https://www.qcif.edu.au/keypoint/.

Parker, T. 2017. "The DataLab of the Australian Bureau of Statistics." *Australian Economic Review* 50, no. 4: 478–483. https://doi.org/10.1111/1467-8462.12246.

Perry, A., and S. Netscher. 2022. "Measuring the Time Spent on Data Curation." *Journal of Documentation* 78, no. 7: 282–304.

Productivity Commission. 2023. "5-Year Productivity Inquiry: Advancing Prosperity." Inquiry Report, no. 100, Canberra.

Scheibner, J., N. Kroesche, L. Wakefield, T. Cockburn, S. M. McPhail, and B. Richards. 2023. "Does Legislation Impede Data Sharing in Australia Across Institutions and Jurisdictions? A Scoping Review." *Journal of Medical Systems* 47: 116. https://doi.org/10.1007/s10916-023-02009-z.

SeRP Australia. "Secure eResearch Platform." Accessed November 9, 2023. https://serp.ac.uk/serp-australia/; Secure Unified Research Environment. Sax Institute. Accessed November 9, 2023. https://www.saxinstitute.org.au/solutions/sure/.

Taylor, L., R. Schroeder, and E. Meyer. 2014. "Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?" *Big Data & Society*: 1–10. https://doi.org/10.1177/2053951714536877.

Varian, H. R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28, no. 2: 3–28. https://doi.org/10.1257/jep.28.2.3.

Veldkamp, L., and C. Chung. 2024. "Data and the Aggregate Economy." *Journal of Economic Literature* 62, no. 2: 458–484.

Villalobos, P., A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. 2024. "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data", ArXiv [cs.LG], arXiv. https://arxiv.org/abs/2211.04325.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 15, no. 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Appendix

What Is a Virtual Desktop?

A virtual desktop is a software emulation of a hardware device that runs on a physical or virtual machine at a remote location, hosted either on premises or in the cloud. In the case of the MIDL, a virtual desktop runs on an on-premises cluster hosted in Melbourne, Victoria. Compute resources (CPU, memory, and storage) are shared across hundreds of virtual desktops that are started and stopped depending on demand. A series of information security controls is in place to ensure that only approved data sets are accessible from a project, and project-level data is only visible from inside the project. The MIDL emulates a Windows 10 environment (see Figure A1) that hosts a range of statistical packages needed for data analysis.

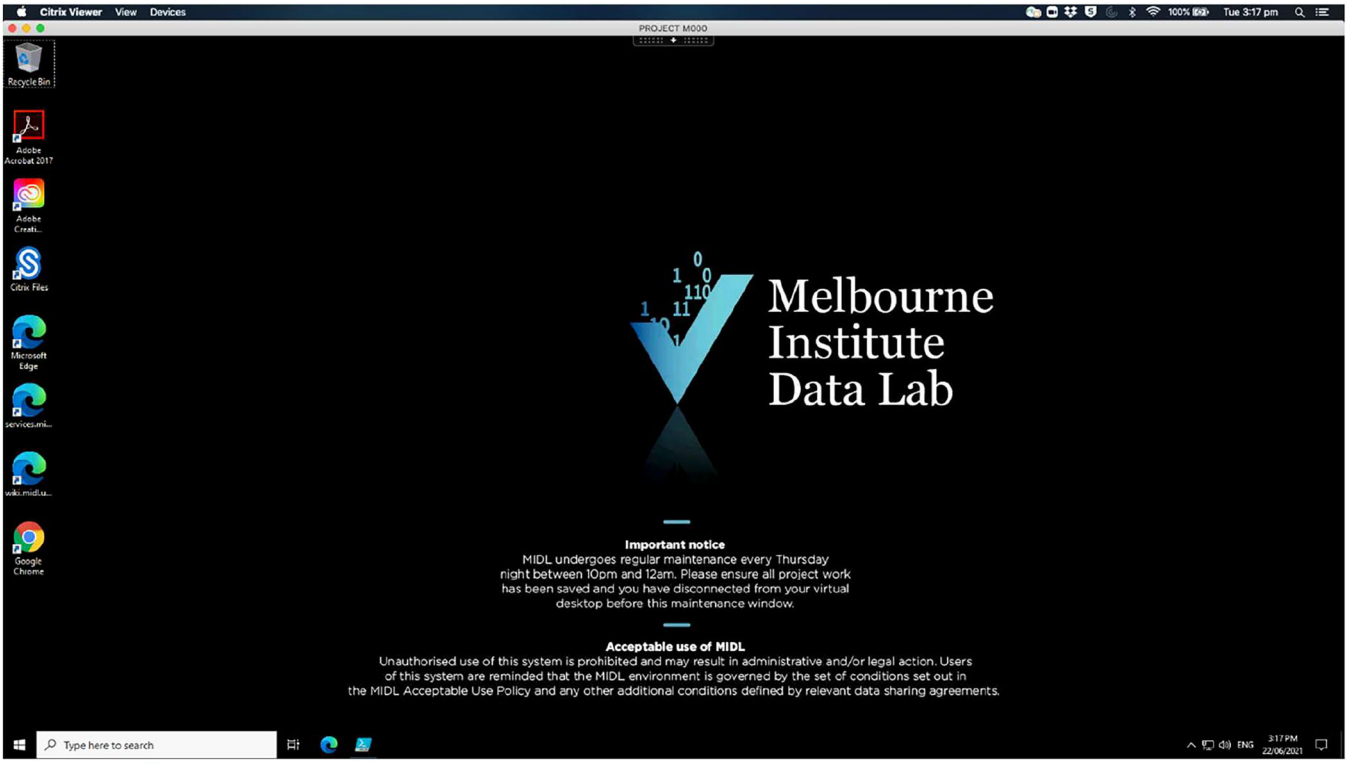

FIGURE A1 | View of a virtual desktop deployed inside the MIDL.