Technical Report

WP4 Demonstrators – D2. Sensitive Environments, Integrated Research Infrastructure for Social Science (IRISS)

Melbourne Institute: Applied Economic & Social Research, The University of Melbourne

December 2022

Table of Contents

Purpose

The purpose of this document is to provide an overview of the Technical Solution for Work Package 4, Part 2 of the Integrated Research Infrastructure for Social Science (IRISS) project as funded by the Australian Research Data Commons (ARDC). This section in the work package relates to assessing the suitability of outputs of IRISS for use in sensitive data environments.

1 Overview of IRISS

The Integrated Research Infrastructure for Social Science (IRISS) project is a project led by the Australian Data Archive at the Australian National University with partners at the Institute for Social Science Research at the University of Queensland, the Australian Urban Research Infrastructure Network, the Melbourne Institute: Applied Economic & Social Research at The University of Melbourne and Australian Research Data Commons. This project intends to address the existing fragmentation between research and the existing research infrastructures used for data integration, research, analyses, and archiving. IRISS aims to establish a new foundation for integration of data, analysis, and platforms for social science research in Australia.

1.1 Overview of Work Package 4

Work package 4 aims to conduct three small scale demonstration projects to establish and illustrate the services provided by the IRISS project. These projects are (1) a spatial data analysis demonstrator that will establish a new data product that leverages Census data outputs and the Geosocial service in IRISS; (2) a sensitive data analysis demonstrator that will assess the outputs of IRISS services to assess their suitability for use in sensitive data environments; and (3) the Australian Census Digital Collection demonstrator that will establish a new pilot collection that will create machine readable data assets and documentation from the Australian Census data collection.

The focus of this document is on part (2) of this work package.

1.2 Core aim of WP4 D2

To assess the outputs of IRISS services such as new surveys and curation tools generated by the IRISS project for use in sensitive data environments.

2 Importance of data sensitivity and environments for sensitive data

Recent years saw an increase in the use of administrative & privately collected business datasets, i.e., the data collected by government agencies & businesses for the purposes of their operations and not specifically for research. While traditionally typical data used for social science research was *made data*, data which is collected through experimental methods, and were designed and collected to address well defined hypotheses[1], the administrative & business data is collected routinely to support the mission of a government agency or an operation of business entity/NGO. Such datasets offer important insights into the efficiency of government policies as well as determinants of individual behaviour. Prominent examples of usefulness of such data include tracking the impact of COVID-19

[1] Connelly et al. (2016) The role of administrative data in the big data revolution in social science research, Social Science Research, 59, 1—12.

and the government response using private sector data[2], combating economic disadvantage[3]. In Australia, administrative datasets have been used to study tax compliance[4] and matching superannuation contributions[5].

Such datasets are large and complex and require special means of handling, cleaning, and management. Because these data were not collected for the purpose of research it might have complex structures or highly unstructured components. As such a typical exercise to clean the data for research, undertake appropriate de-identification of unit record data and other practices used in statistical disclosure control[6], may have not been applied in many cases. It is also common for the researchers to link them with other types of data, often geographical or temporal. The linking of the data from different resources brings additional challenges and risks.

A related development is increased use of microdata derived from a population Census. Integrated Public Use Microdata Series (IPUMS) developed by the Institute for Social Research and Data Innovation in the University of Minnesota was one of the first popular data product developed using individual-level data derived from the population censuses. In Australia, Australian Bureau of Statistics developed a set of products using individual data from Australian Census, such as Australian Census Longitudinal Dataset (ACLD), Business Longitudinal Analysis Data Environment (BLADE), and others.

The increased use of such datasets for research and policy purposes brough a rise in awareness of a set of legal and ethical issues. The use of such datasets for academic and policy research requires a careful consideration of privacy and confidentiality risks. This is especially true when the data contains sensitive information about individuals, such as health records or financial information. Data security becomes an important domestic and international concern for governments and private organisations. Recent data breaches of Optus and Medibank serve as reminders for the importance of data security.

All these factors lead to the importance of the appropriate research infrastructure that enables research on pertinent topics while also protecting sensitive data from unauthorised use and malicious actors. The research infrastructure should be designed to ensure that data is secure and protected from unauthorised access. This includes the use of encryption, authentication, and access control measures.

For privacy and confidentiality reasons access to sensitive data, especially in the case of administrative data sources, is tightly controlled by governing bodies and/or the data custodians. Typically, this is done by having (1) data access approvals processes; (2) researcher requirements such as additional training and (3) increased physical/virtual security settings. Traditionally some of the requirements for (3) were imposed using physical purpose-built rooms where researchers used to access specific data and were unable to take any data outside these rooms. Nowadays with advances to network infrastructure and cloud computing, a range of information security controls can be imposed on your environment and tailored at a user or project-level basis.

---

[2] Chetty, R., Friedman, J. N., Hendren, N., & Stepner, M. (2020). Real-Time Economics: A New Public Platform to Analyze the Impacts of COVID-19 and Macroeconomic Policies Using Private Sector Data, Working paper

[3] Chetty, R. (2021). Improving equality of opportunity: New insights from big data. *Contemporary Economic Policy*, *39*(1), 7—41.

[4] Lim, Y., Evans, C., & Kayis-Kumar, A. (2022). The exploitation of tax professional expenses for tax minimisation: evidence from Australia. In *Australian Tax Forum*, *37*(2), 295—318.

[5] Polidano, C., Chan, M., Vu, H., Wilkins, R., Carter, A., & To, H. (2020). How Effective are Matching Schemes in Enticing Low-income Earners to Save More for Retirement?, Working paper

[6] Statistical disclosure control (SDC) is a technique used in data-driven research to ensure no person or organization is identifiable from the results of an analysis of survey or administrative data, or in the release of microdata. The purpose of SDC is to protect the confidentiality of the respondents and subjects of the research.

3 User requirements

Table 1 provides the User Requirements for the D2 Work Package that were developed in June 2022. They are provided here to provide context and serve as input to the technical solution that is proposed in this report.

Table 1. User Requirements for Work Package 4, D2 - Sensitive Environments.

URS ID	User Requirement
1. Access to environment	
1A	Users shall have an easy process (user journey) to follow to gain access to the sensitive data environment.
1B	Users shall have the ability to test the sensitive data environment prior to creating a production-ready process.
1C	Users shall have a process by which additional requirements may be imposed on researchers intending to access relevant data assets/tools that they create for other approved users within the sensitive data environment.
2. Technical requirements for appropriate function	
2A	Users shall have access to appropriate technical resources and technical support to undertake any custom work (additional fees may apply) that may need to be undertaken for a project.
2B	Users shall have access to required websites, API's that may be needed for the appropriate function of their tools within the sensitive data environment.
2C	Users shall have access to file ingress and egress capability to bring relevant files in to and out of the sensitive data environment.
2D	Users shall have access to host their data asset/tools for other approved users within the sensitive data environment.
2E	Users shall have access to host their data asset/tools for other approved users outside the sensitive data environment.

4 Solution overview

The Work Package 4, D2 Sensitive Environment demonstrator project will be demonstrated within the Melbourne Institute Data Lab (MIDL), a secure data enclave within the Melbourne Institute at The University of Melbourne. It is important to note that many aspects of the other demonstrators within the IRISS project will be developed, tested, and deployed outside the MIDL secure environment. Sensitivity of data used for demonstration purposes, development and testing regimes allow the other demonstrators to be developed in less secure settings. The scope of WP4, D2 is to allow these demonstrators to be migrated and/or accessed from within the MIDL's secure environment – in line with the core aim set out in Section 1.

4.1 What is MIDL?

MIDL is a secure, purpose-built data enclave that enables virtual access to micro-level data for curation, analysis, and visualisation. It is a collaboration between the Melbourne Institute: Applied Economic & Social Research, a research department within the Faculty of Business and Economics at the University of Melbourne, with a mission to inform Australian economic and social policy; and Cyconsol, an Australian-based professional services provider specialising in the design and operation of cybersecurity and ICT capabilities working with Australian Government and industry.

4.1.1 Service offering

The key services of MIDL include:

- A virtual computing environment with an increased security posture to host a range of sensitive data assets to approved users for the purpose of research that informs Australian economic and social policy. Data sets may include assets from the Australian Federal Government, state and local governments, service providers and industry. Data sets with security classification up to PROTECTED (or similar in cases where data custodians are outside Government) may be stored and hosted for approved users through MIDL.
- Safe and secure processes to protect data assets of data custodians minimising risk of disclosure.
- Access to a secure information management system (MIDL Wiki) from inside the environment which provides an increased value proposition for both researchers and data custodians. This Wiki page provides users with additional data documentation, "research ready" and a "working data set"[7] versions of data assets as curated by research staff at the Melbourne Institute enabling faster research.
- Ability for researchers to use/combine a range of data sets for analyses fitting a specific research theme through a Shared Data Environment (SDE). The purpose of a SDE is to permit researchers interested in working on a given theme to have access to data from a range of sources, with permission, that allows better analyses of questions related to the theme. SDE users have the additional capability to contribute to the body of knowledge through the MIDL Wiki system.
- Access to additional support, data services and engaged research opportunities using Melbourne Institute research and professional staff.

4.1.2 Security and regulatory compliance

MIDL has undergone a security assessment under the Australian Government's Information Security Registered Assessors Program (IRAP) for a security classification of PROTECTED. The MIDL team is currently targeting to have MIDL undergo accreditation to be an Accredited Data Service Provider under the recent Data Availability and Transparency Act (2022).

4.2 Technical design

The MIDL's design is split across two components: infrastructure – component made up of shared resources for MIDL that are not directly associated with an end-user, project, or shared data

[7] A "working data set" is a data set or collection of data sets that have undergone further transformation to create a single, harmonized data asset that is ready for analysis of a specific research theme or topic. For more information, please see Payne and Samarage (2022) Maximising evidence-based policy analysis through data sharing *in* Dawkins and Payne (Eds) (2022) Melbourne Institute Compendium 2022: Economic & Social Policy: Towards Evidence-Based Policy Solutions. Melbourne Institute: Applied Economic & Social Research, University of Melbourne, Australia.

environment; and workload – component made up of various workloads (desktops) hosted within MIDL to enable research outcomes.

The MIDL's virtual desktop infrastructure primarily leverages hyperconverged infrastructure, secure network protocols and a virtualisation infrastructure to provide researchers with the ability to securely access and conduct data analysis on both Government and non-Government data. The MIDL is hosted on a hyperconverged Nutanix infrastructure that allows for seamless sharing of a large pool of CPU compute and memory resources. This allows the MIDL to virtualise users' access to sensitive data (as required or as approved for access) and relevant software applications needed for data analysis.

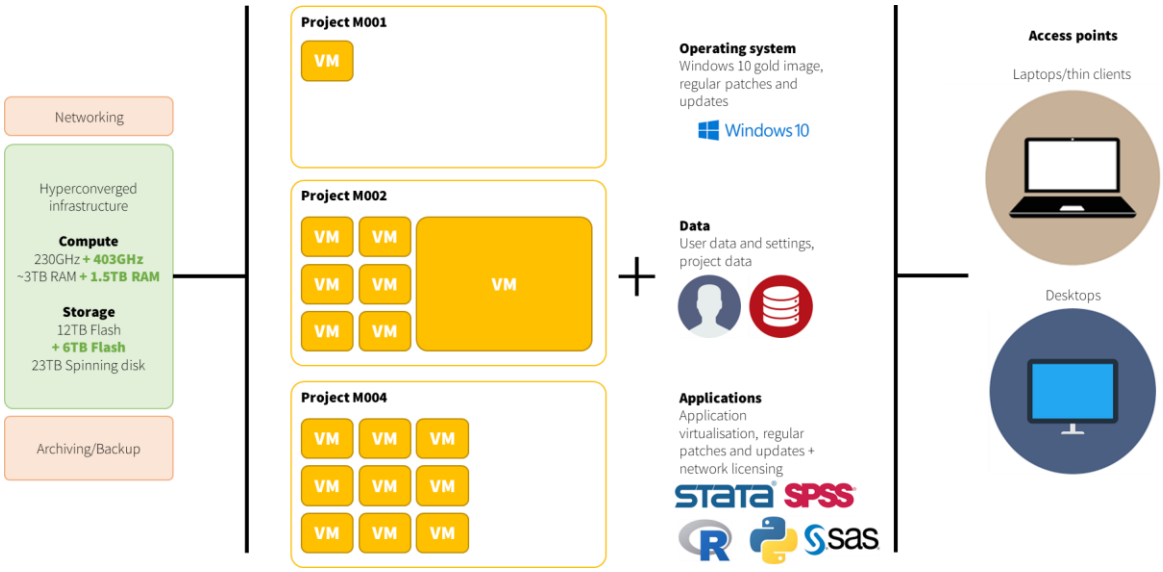

Figure 1. MIDL's hyperconverged infrastructure

The MIDL's network infrastructure including servers are contained within a SCEC Class C rack and housed at a physical data centre with sufficient physical security controls needed for a PROTECTED platform. All systems and services are located and managed from within Australia.

The MIDL also enables secure (encrypted) back up and archiving of data both on hot and cold storage for a period of up to 20 years. The MIDL's environment and facilities are also monitored 24/7 by a third-part security incident event monitoring service resourced with a full security operations team.

All users must undergo annual security awareness training and vetting (privileged users only) before being granted to MIDL's systems. For more information, please refer Section 4.4.

4.3 Operational design

MIDL's secure virtual desktop infrastructure (VDI) allows approved users to remotely access and analyse sensitive data assets from primary and secondary sources for approved studies. MIDL's VDI is setup as follows:

- A **project workspace** is defined by the research topic/question set out in the MIDL Project Application Form. All users named in the application form will be approved for a project on MIDL by the data custodian(s) for the requested data. These approved users will have access to the project workspace on MIDL. An illustration of this provided in Figure 2.

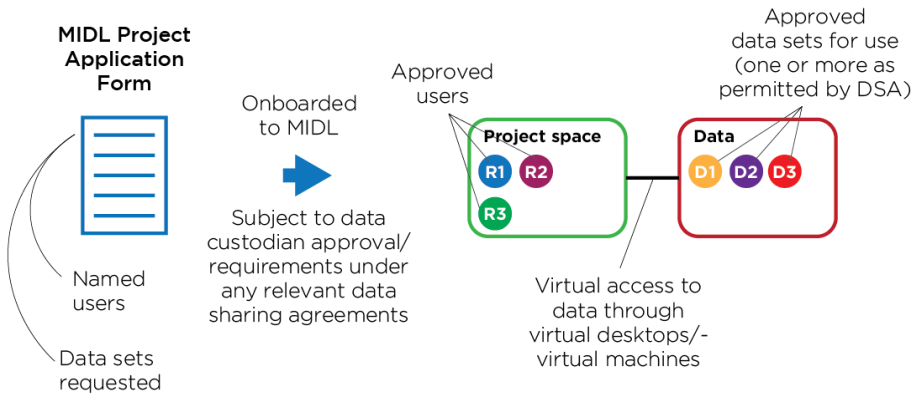

Figure 2. Project workspace within MIDL.

- Each project workspace has a number of **virtual desktops** that they may utilise to access the computing environment for analyses. These virtual computing environments are accessible through these virtual desktops over an encrypted internet connection and provide users with a range of statistical analytical packages and programming languages for data analysis and visualisation. MIDL is committed to enable the use of flexible workspaces i.e. the use of virtual desktops across multiple users in a project providing an economical solution to researchers.

The data sets that are requested for use through the MIDL Project Application Form will need to be approved by data custodian(s) prior to access through the virtual desktops mentioned above. In MIDL, users can access data through two mechanisms:

1. through a **standalone project** workspace which is created for a research study that is approved by data custodians for investigating a specific set of research questions within a singular theme, or
2. through an already defined **Shared Data Environment** (SDE), a special project workspace specifically set up for projects that conduct analyses and research on a specific research theme.

4.3.1 Standalone projects

A standalone project is defined by a singular research topic as approved by the data custodian(s) using the information provided on the MIDL Project Application Form. Standalone projects consist of a single project workspace through which approved users can access a single data set or multiple data sets (subject to requirements from the data custodians) for analyses. If there are multiple data sets, approved users may combine these data sets to enable deeper analyses. The use of multiple data sets may not be available on all projects and are subject to requirements defined in relevant Data Sharing Agreements (DSA) between the University of Melbourne/ Melbourne Institute and the data custodians.

4.3.2 Shared Data Environments

A Shared Data Environment (or SDE) is a special project workspace that is set up for projects that conduct evidence-based research to inform Australian economic and social policy under a singular theme e.g. study of disadvantage and cycles of disadvantage in Australia. SDE's are first developed by a team of researchers interested in a collaborative research space for a specific theme of research. Multiple data sets may be brought in to the SDE to enable research under this theme across multiple projects. A data governance committee will be established by these lead researchers to enable a regulated process in bringing additional data in to the SDE.

SDE's provide approved users with a range of other features that may not be available through standalone projects. These features are enabled through contributions to a shared body of knowledge inside the SDE by MI staff and other approved users inside the SDE. These include the following:

- Memoranda, codebooks, and other materials to assist the users in their understanding of the data for which they have been given access.
- Access to a sub-sample data set (what we refer to as a "research ready" and/or "working data set") that contains the popular variables amongst similar researchers.
  e.g. projects that look into income will have the relevant income variables in their extract. Projects that are investigating economic disadvantage will include additional information on the income variables and treatments applied to them because of how the data were collected.
- Access to programs used to transform and create sub-samples of data sets including the above mentioned "research ready" versions of a data set.
- Access to memoranda, sample code and derived variables other approved users (within the same SDE) that have been shared from prior analysis. Users in the SDE may also contribute to this shared body of knowledge which is housed inside MIDL's information management system (or MIDL Wiki).

This collaborative environment will be enabled through the MIDL Wiki page that will only be accessible for approved projects from within MIDL. If your project workspace is part of a Shared Data Environment (SDE), this feature will be available by default unless instructed by the relevant data custodians or data sharing agreement.

For more information on Projects and how they differ to Shared Data Environments, we refer the interested reader to Part 1 of the MIDL Information Pack[8].

[8] Part 1 of the MIDL Information Pack titled "What is MIDL?" can be accessed from this link.

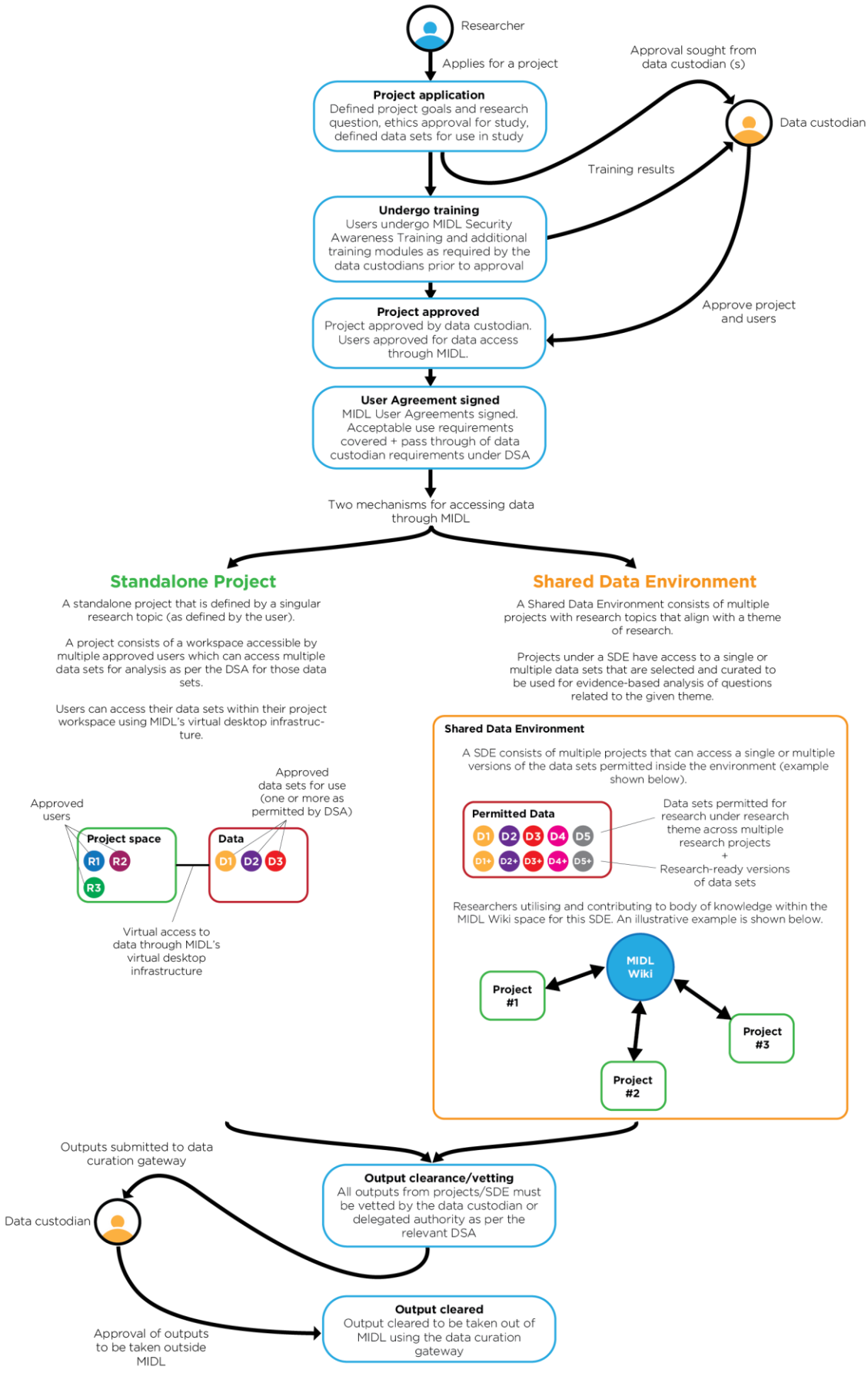

Figure 3. MIDL user journey and mechanisms of data access.

4.4 User journey

The MIDL's user journey to create a new project or onboard new users have been designed with the Five Safes Framework in mind. Shown below is an overview of the MIDL User Journey which is aimed to address user requirement 1A set out in Table 1. The interested reader is directed to Part 2 of the MIDL Information Pack for more information[9].

Figure 3 shows a graphical representation of the MIDL user journey. Key aspects of the user journey include:

- The requirement for all users (and privileged users) to undergo MIDL Security Awareness Training prior to being granted access to MIDL's systems.
- The process by which data custodians can be involved in reviewing a user request to access their data assets hosted within the MIDL before being granted access to data.
- There is a legal mechanism in place within the MIDL's user agreement (the MIDL Access Agreement) in which data custodians can stipulate additional requirements for a specific project and its membership of approved users.
- Another key part of the user journey is the ability to take out completed outputs from the MIDL environment. No complete data assets can be taken outside the environment. Depending on restrictions set out by Data Sharing Agreements, users may need to have their outputs vetted by the data custodian or a MIDL output vetting specialist before they can be released outside the secure environment.

Figure 4 provides a graphical representation of an example scenario where a data user (researcher) is accessing their project workspace inside the MIDL. Through their virtual desktop (VM1), they can access data made available by the data custodian as well as additional material made available by the Melbourne Institute Data & Analytics team. These additional materials include a "research ready" data asset as well as additional documentation about the data set including program code to replicate the production of the research ready data asset. This figure also includes a flow for output vetting and clearance by the data custodian before release to the data user outside of the MIDL.

[9] Part 2 of the MIDL Information Pack titled "MIDL for Users" can be accessed from this link.

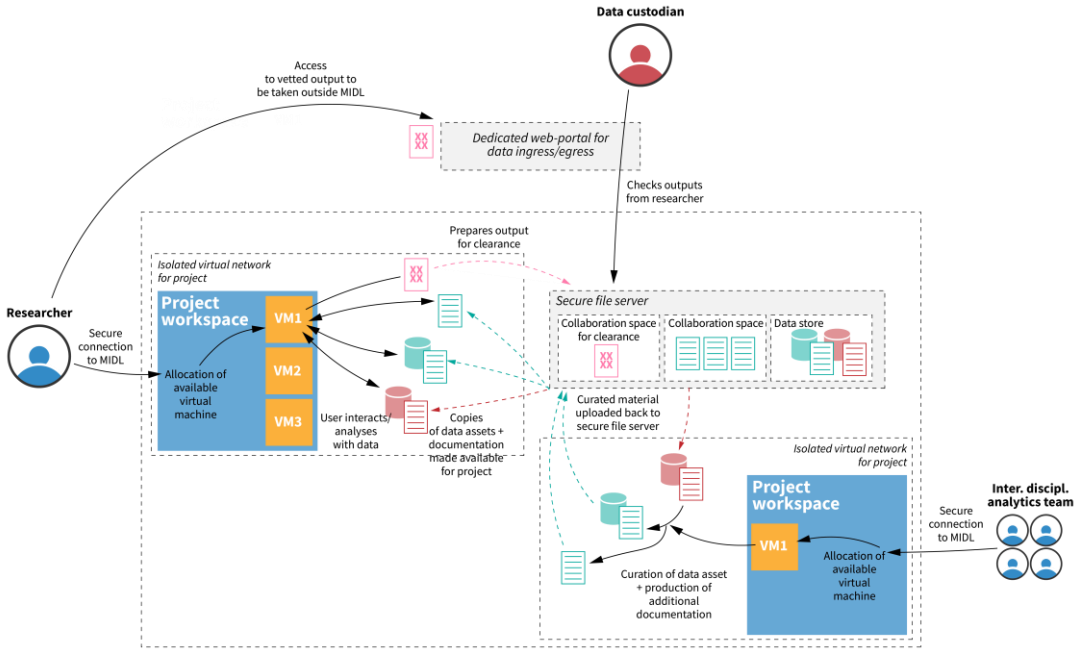

Figure 4. Data access workflows - an example scenario

4.5 Adherence to the Five Safes Framework

This section covered aspects of the MIDL's design at a high level. This design was intended to work alongside the Five Safes Framework, (or also referred to as the Data Sharing Principles), a framework used by several Government agencies and Data Custodians in Australia as well as multiple organisations overseas. The Five Safes Framework is applied across five distinct domains: projects, people, data, settings (what is determined to the infrastructure used for data access) and outputs (from this infrastructure). Table 2 provides an overview of how the MIDL's technical and operation design adheres to the Five Safes Framework.

Table 2. Application of the Five Safes Framework within MIDL.

Element of framework	Application through MIDL
Safe People	Users must undergo security awareness training and complete an authorisation process prior to data access. The organisation(s) that potential users belong to must first sign MIDL Access Agreements that form a head agreement between the organisation and the University of Melbourne for 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 prior to being granted MIDL access. This project schedule includes a confidentiality agreement that stipulates a user's requirements to maintain data privacy and confidentiality whilst using MIDL. The MIDL security awareness training covers shared responsibilities of user's in maintaining data security and data confidentiality of assets held within the MIDL. It also includes key contact details for raising service requests and security incidents. 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 nominate a user approval process that requires additional security controls such as police checks and/or security clearances prior to granting access to data.
Safe Projects	Current processes requires 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 would also ensure that project aims align with a data custodian's mission statement and values.
Safe Settings	Information security controls implemented within MIDL for a security classification of PROTECTED under Australian Government Protective Security Policy Framework and the Information Security Manual (ISM). MIDL utilises an array of information security controls as required by Australian Government and international regulations. This includes: • strong authentication and identity management controls including the enforcement of multi-factor authentication protocols • 24x7 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 in to or out of the MIDL environment In mid-2022, MIDL completed its first security assessment under the Australia Government's Information Security Registered Assessor's Program (IRAP) by an external auditor. MIDL was assessed against the information security controls required for a PROTECTED-level system under the ISM and found to be 93% compliant against the assessed controls. MIDL also underwent a penetration test and will undergo annual penetration testing and further IRAP assessments every two years as per MIDL's assurance and audit policies.
Safe Data	MIDL accepts data that have direct identifiers removed with further treatments applied to minimise disclosure risk. Where this is not applicable, MIDL will leverage its increased security posture with additional security checks for users that need access to sensitive information. Data containing direct identifiers will not be released to the user from inside the secure environment.

	A range of statistical disclosure control tactics are highlighted in the MIDL security awareness training as techniques that users can implement on their outputs prior to preparing for vetting by the data custodian.
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 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, limited internet functionality to ensure data cannot be taken out without the use of the MIDL output clearance processes.

4.6 Governance

Within the MIDL platform, 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 (in terms of IRAP requirements with future scoping of work to include ISO);
2. data; and
3. governance templates for specialised projects, the Shared Data Environments (SDE) hosted within MIDL.

Through the MIDL Governance Framework, we have several governance structures to ensure appropriate oversight, roles and responsibilities and approval mechanisms are in place to achieve the above. Data management within the MIDL is overseen by the MIDL Data Governance Body. The MIDL Data Governance Body oversees the data onboarding and access approval processes for all projects/users within the MIDL with support from the MIDL Services Team. This includes communications processes to set up with data custodians on user/project approvals, execution of legal agreements for MIDL access, review of training outcomes, data onboarding processes, output vetting requirements and additional requirements that may be placed by the data custodians. MIDL's Data Governance (Figure 5) is achieved through the Data Governance Body and a range of expertise roles such as Data Stewards, Data Specialists, Data Curators, Data Users (researchers) and Output Vetting Specialists.

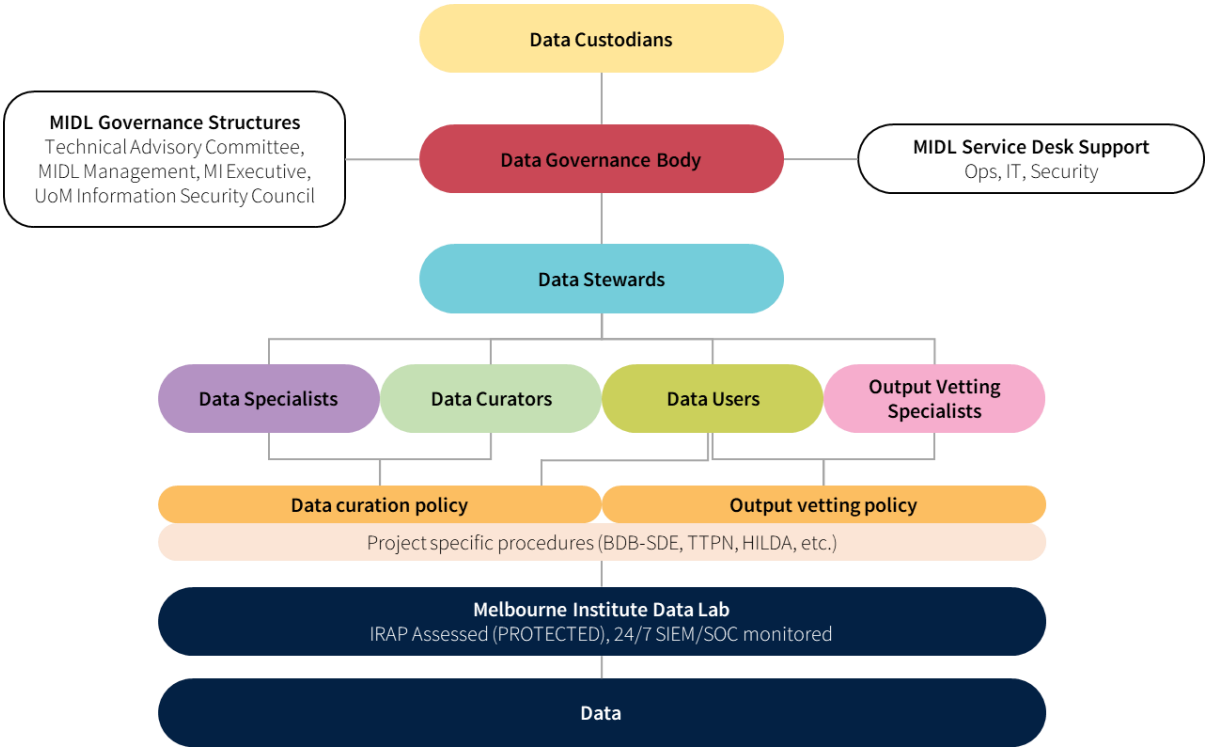

Figure 5. MIDL Data Governance Framework

5 Integration with IRISS

5.1 Integration with WP4-D1 GeoSocial Demonstrator

5.1.1 Overview of the GeoSocial demonstrator

Work Packages 3 and 4-D1 of the IRISS project focusses on developing and piloting a "proof of concept" data integration service called *GeoSocial*. GeoSocial allows researchers to enhance people-centred survey data such as HILDA, LSAY, LSAC or TTPN, with spatially structured data which includes information on places where these individuals live.

The figure below provides a lower-level use case diagram for the GeoSocial service. It consists of the following key features of the GeoSocial service:

1. Data inputs including attributes provided by the researcher;
2. The core data integration solution that draws on data inputs and utilises a script repository to undertake the data integration. This script repository that consists of scripts to perform data integration, which are implemented using R;
3. The HTML-based reference page providing background information about the GeoSocial service.

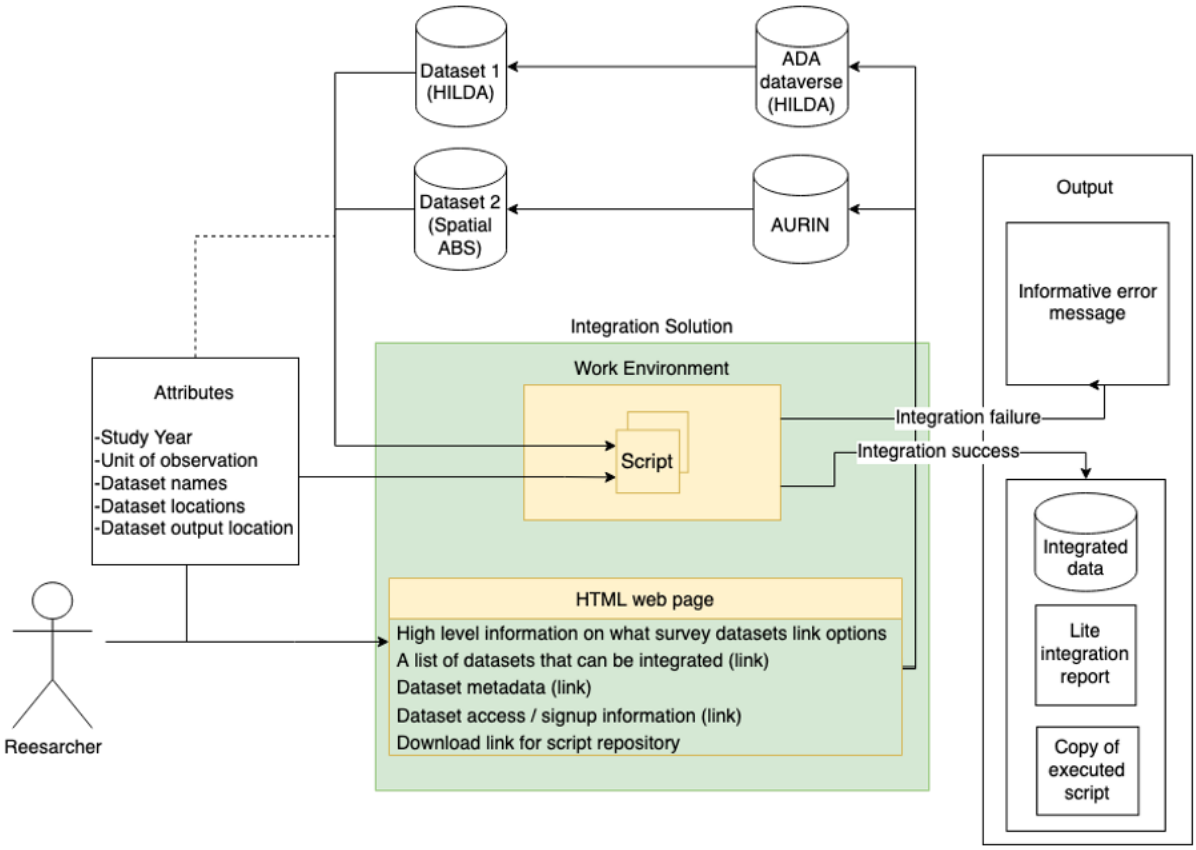

Figure 6. Use case diagram for the GeoSocial service

5.1.2 Implementation within the MIDL

Piloting of the GeoSocial service within MIDL under Work Package 4 – D2 will be implemented through migration and reconfiguring. This activity will commence by setting up a *project* within the MIDL to house this activity. A solution architecture diagram for this implementation is shown in Figure 7 – which is a modified version of the image shown in Figure 6.

The GeoSocial piloting project will allow the IRISS team to assess how they would implement a data integration exercise such as that done with GeoSocial from within an environment with tight security controls. In line with user requirement 1B, the use of a project structure enables the IRISS team to test data integration scripts inside a secure environment as part of the development strategy.

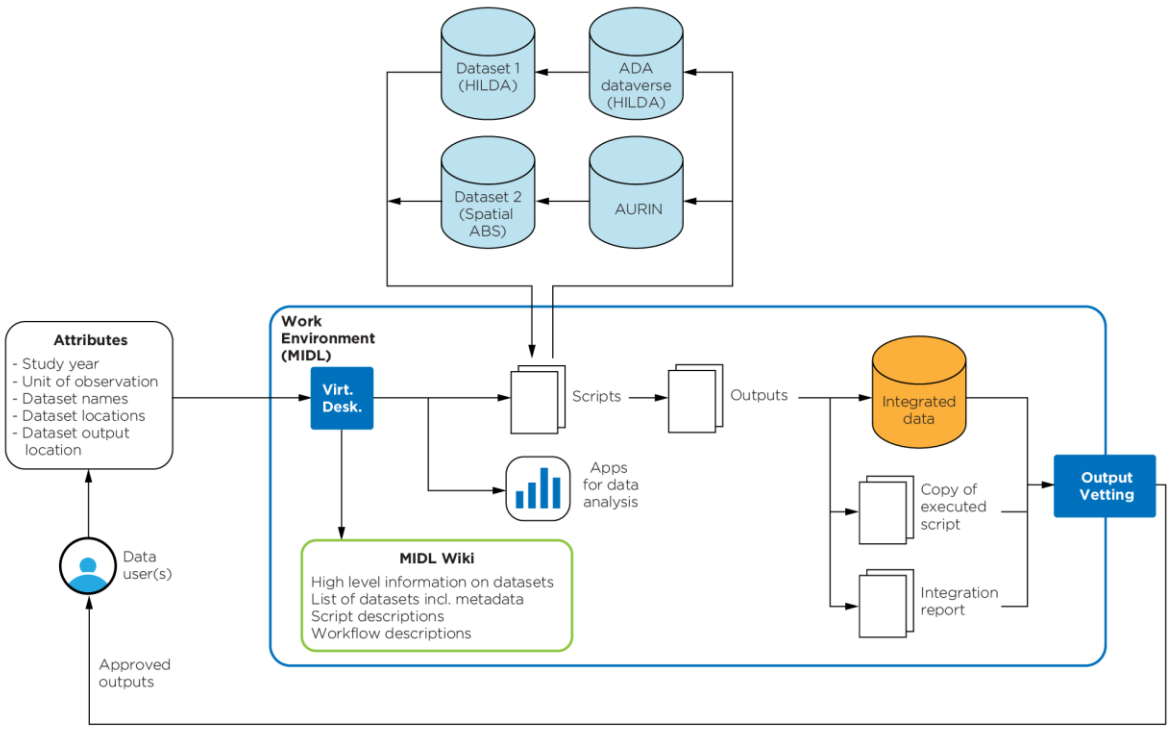

Figure 7. Implementation of GeoSocial (IRISS WP3/WP4-D1) within the MIDL secure research environment.

Data users can access the GeoSocial project materials from within a dedicated virtual desktop that virtualises a Windows 10 environment.

The GeoSocial documentation including data descriptions, scripts and accompanying information on relevant short descriptions of the scripts, its workflow and rationale behind key steps would be contained within a dedicated wiki space (collection of pages and content) inside the MIDL Wiki. The MIDL Wiki will also contain usage instructions for the scripts with example scripts for researchers to run the data integration activity. This ensures that researchers from a range of analytical expertise can effectively undertake the data integration activity.

This virtual environment will contain a web shortcut to access the HTML web site proposed in the GeoSocial work package. This will be enabled by whitelisting the relevant URL such that the web content can be accessed from within the MIDL environment. Alternatively, provisions can be made (as per user requirements 2A and 2D) for the HTML web site to be directly hosted to the public within the MIDL's AWS de-militarised zones.

For the data integration scripts to work, specific firewall rules will be enabled to ensure the script can communicate with servers at the ADA and AURIN. This is in line with user requirement 2B.

Once the activity has been completed, the GeoSocial package creates a copy of the script that was executed along with the integrated data asset (depicted in orange as we would assume that data that has been integrated is more sensitive than its parts). Specific output vetting rules and procedures will be created in line with the MIDL's Output Vetting Policy to ensure that integrated data cannot be taken out of the MIDL environment. Researchers wishing to take outputs out of the environment will need to utilise the MIDL's output vetting and clearance features to have their outputs vetted before release. This is in line with user requirement 2C.

5.2 Integration with WP4-D3 Census Demonstrator

5.2.1 Overview of the Census Demonstrator

The use of Census data forms the foundation of much major data-oriented research in Australia, and represents the key reference data asset in Australia on human population and settlements. Access to the aggregate or micro-level Census is available through multiple services by the Australian Bureau of Statistics (Table Builder, ABS DataLab, ABS Data Services). But this is mainly to the most recent Census and as far back as 2001. Access to historical Census data is more challenging as with a significant proportion of population data is in archives and libraries or digitised but embedded in formats largely inaccessible to researchers or to the Australian public.

Discussions are currently ongoing surrounding the scope of this demonstrator and the work that is to be carried out. Initial conversations are around supporting a piece of work that is to be undertaken by the Historical Frontier Violence project (HFV). The HFV project, led by the Melbourne Institute, aims to integrate massacre data dating back to pre-colonisation times, geographical data including weather, rainfall, soil quality, historical population data starting from the first Census to understanding the impacts on communities in the present day.

5.2.2 Implementation within the MIDL

The implementation of this work package within the MIDL will be driven by the exact scope of work set out in the Census Demonstrator. Due to our limited understanding of the scope of this work, the outline provided here is preliminary and would be further fleshed out in future reports.

This activity will be implemented as a *project* within the MIDL environment. Preparation of Historical Census data extracts will be undertaken by the Australian Data Archive (ADA) team on ADA systems. We hope to mimic the data integration activities undertaken by the ADA inside the MIDL but this work may be outside the scope of the IRISS project in its current phase.

Additionally, we will create a dedicated wiki space for the Historical Census demonstrator that captures information on the integrated data set, variables integrated, and method of integration used. The wiki space could also host data quality statements on the integrated Census data such that researchers intending to use these data can access relevant documentation from within the research environment.

6 Closing remarks

This first technical report provided an overview of the Melbourne Institute Data Lab (MIDL) platform that is used to test how integrated research activities are carried out within a secure research environment with strict security controls. This report also provided a first view of how the IRISS project's GeoSocial data integration service could be tested within a secure environment; and a preliminary view of how the MIDL could facilitate data integration of historical Census data for understanding historical frontier violence and its impact in today's Indigenous communities.

6.1 Future work

The work set out in this report provide an overview of the work that is to be undertaken in the current phase of the IRISS project. The amount of effort and money expended by the Australian Government to enable better data sharing and access to more secondary data for research purposes, has led to sharp rise in the number of secure research environments to undertake safe data analyses. This calls for better integration and communication between multiple secure sites, something that projects like CADRE (Coordinated Access for Data, Research and Environments), a project led by the Australian Data Archive team at the Australian National University, is aiming to achieve.

Figure 8 shows a graphical representation of a solution within the MIDL that undertakes data integration using both open data (green data cubes), and sensitive data (orange and red data cubes). Red data cubes depict highly sensitive microdata that are only available for researchers through a secure research environment, while orange data cubes depict sensitive data that has undergone integration of two data assets that do not necessarily need a secure research environment for researchers to access.

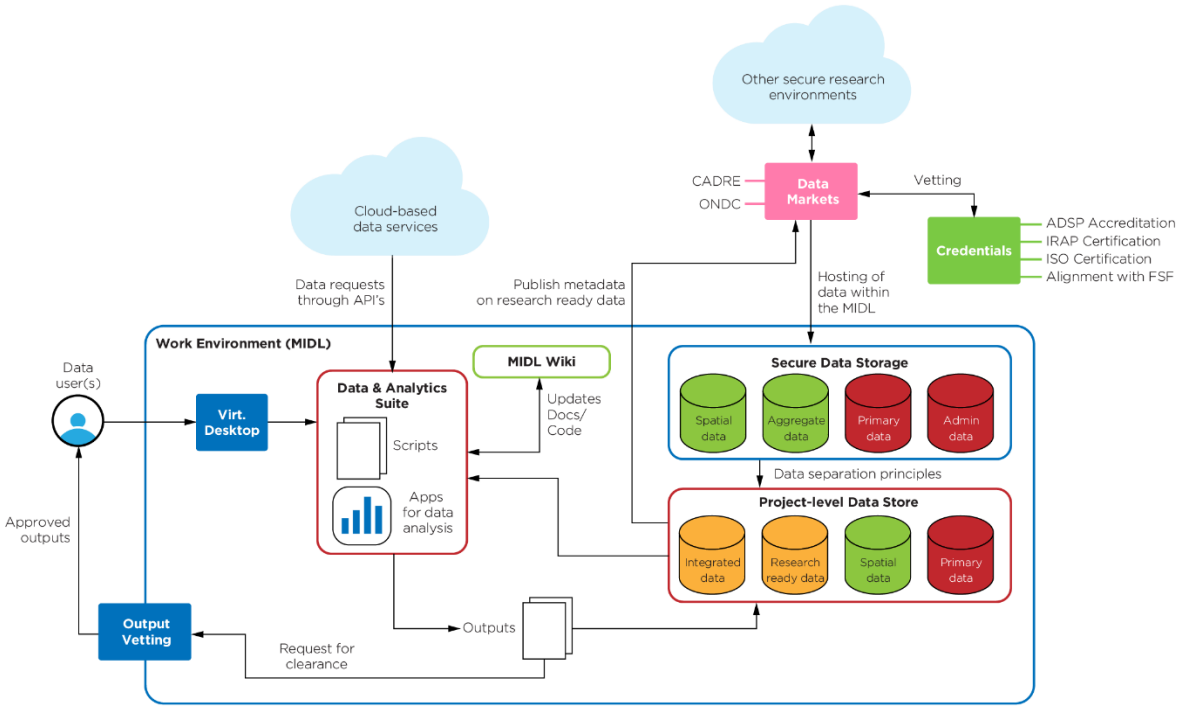

Figure 8. Solution overview for data integration using a range of sensitive data assets and data marketplaces.

This solution outlook sees the researcher (data user) access their secure environment using the same way as today, through a virtual desktop that contains a virtual Windows environment with all relevant statistical and programming packages. Unlike the previous implementation which utilised custom project-level firewall rules for users to access external systems, the use of application program interfaces (API's) to directly communicate and 'pull' data to within the MIDL environment would be a more efficient way to access data.

Furthermore, the MIDL could directly communicate with external 'data markets' that enable data sharing between two entities or gain access to a range of datasets. This includes systems/platforms such as Dataplace (implemented by the Office of the National Data Commissioner), the CADRE platform, the Australian Data Archive's Dataverse systems to name a few. These sensitive data would be duplicated to match an external server such that relevant updates (i.e. new waves, new data, error fixes etc.) are readily duplicated within the MIDL platform. Data separation principles and the use of project-level access controls (currently in place) would ensure that users within a project only have access to data assets that have been previously approved for access within that project.

The above allows researchers to bring data into their research environment from a variety of sources and integrate as needed to undertake deeper and richer data analyses. Specific output vetting rules and procedures in line with the MIDL's Output Vetting Policies can be established to ensure that any data cannot be taken out of the MIDL environment. Researchers wishing to take outputs out of the

environment will need to utilise the MIDL's output vetting and clearance features to have their outputs vetted before release.