Sensitive data integration services using shared data environments

Dr. Rajeev Samarage

Melbourne Institute: Applied Economic & Social Research, The Faculty of Business and Economics
The University of Melbourne

eResearch Australasia 2023 Conference
16 – 20 October 2023, Brisbane, Australia

THE UNIVERSITY OF MELBOURNE

MELBOURNE INSTITUTE
APPLIED ECONOMIC & SOCIAL RESEARCH

Acknowledgement of country

I acknowledge the Traditional Custodians of the land on which we gather today.

I recognise their continued connection to the land and waters of this beautiful place and acknowledge that they never ceded sovereignty.

I pay my respects to the Elders, those who have passed into the dreaming; those here today; those of tomorrow.

> Closeup of bark of a Paperbark tree. Image courtesy of David Clode from Unsplash.

Sensitive data…

Let's start with some definitions from The Privacy Act (1988)

personal information means information or an opinion about an identified individual, or an individual who is reasonably identifiable:

(a) whether the information or opinion is true or not; and

(b) whether the information or opinion is recorded in a material form or not.

- an individual's name, signature, address, phone number or date of birth
- employee record information
- photographs
- internet protocol (IP) addresses
- voice print and facial recognition biometrics (because they collect characteristics that make an individual's voice or face unique)
- location information from a mobile device (because it can reveal user activity patterns and habits)

sensitive information means:

(a) information or an opinion about an individual's:

(i) racial or ethnic origin; or

(ii) political opinions; or

(iii) membership of a political association; or

(iv) religious beliefs or affiliations; or

(v) philosophical beliefs; or

(vi) membership of a professional or trade association; or

(vii) membership of a trade union; or

(viii) sexual orientation or practices; or

(ix) criminal record;

that is also personal information; or

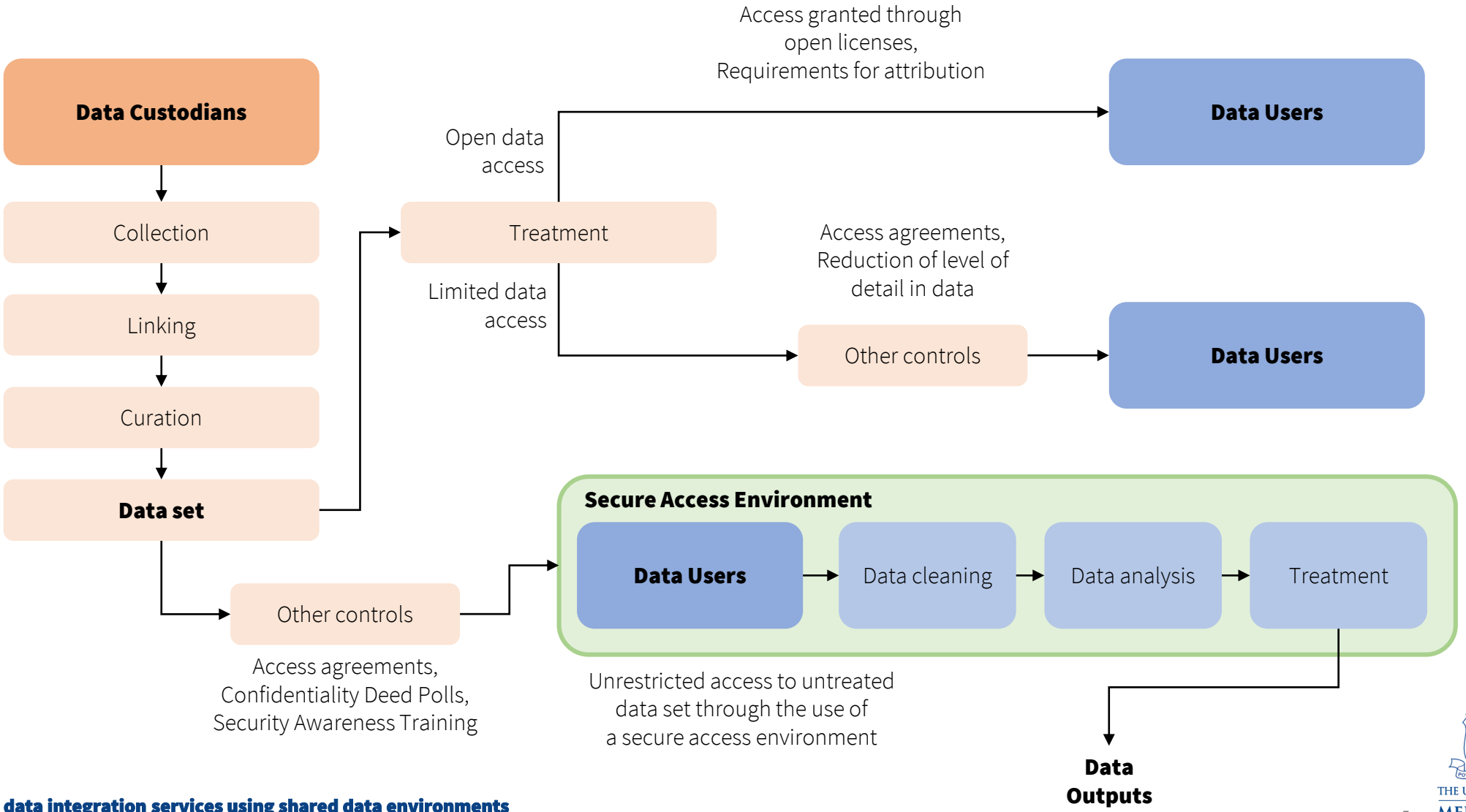
(b) health information about an individual; or

(c) genetic information about an individual that is not otherwise health information; or

(d) biometric information that is to be used for the purpose of automated biometric verification or biometric identification; or

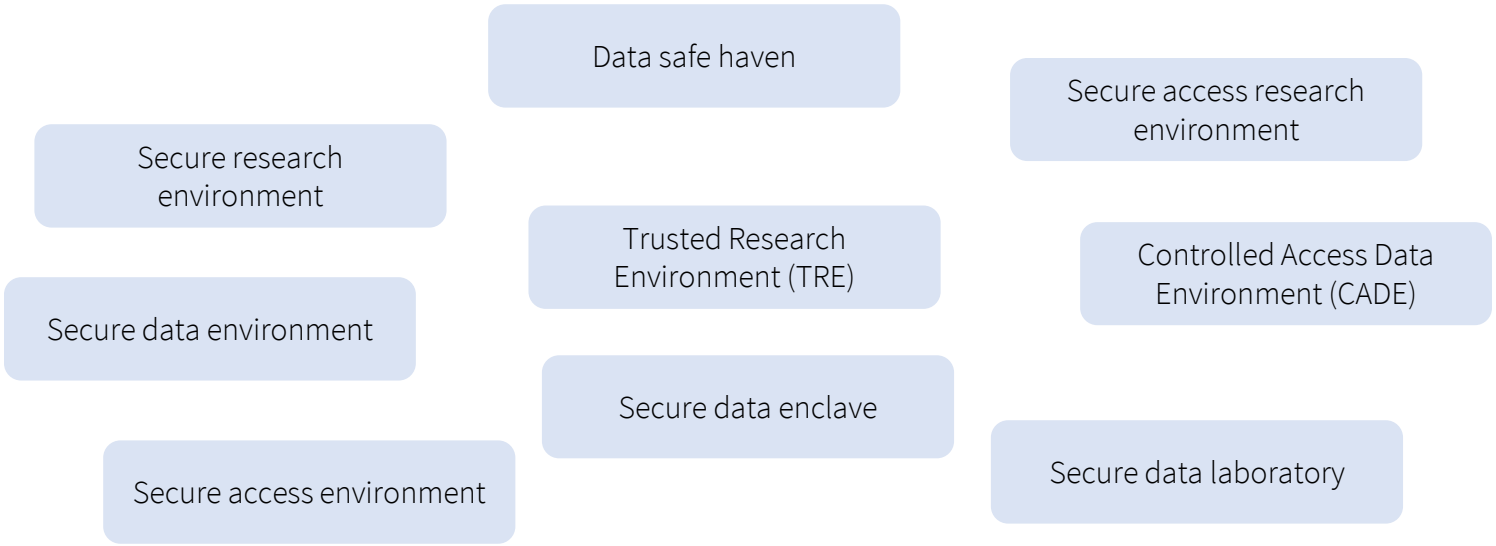
(e) biometric templates.

What has been your experience working with data?



What is a sensitive access environment?

Ongoing debate about terminology



What is IRISS?

- The Integrated Research Infrastructure for Social Science (IRISS) project intends to address **the existing fragmentation** between research and the existing research infrastructure used for data integration, research, analyses, and archiving.
- Starting point for this infrastructure is a core foundation of **data** – its acquisition, documentation, harmonisation and dissemination for re-use.
- Project outcomes:
  - Coordinated data governance and integration
  - Enhanced research infrastructure
  - New data integration environment
  - New researcher training opportunities

Australian Research Data Commons

Australian National University

The University Of Queensland AUSTRALIA

The University of Melbourne

AURIN aurin.org.au

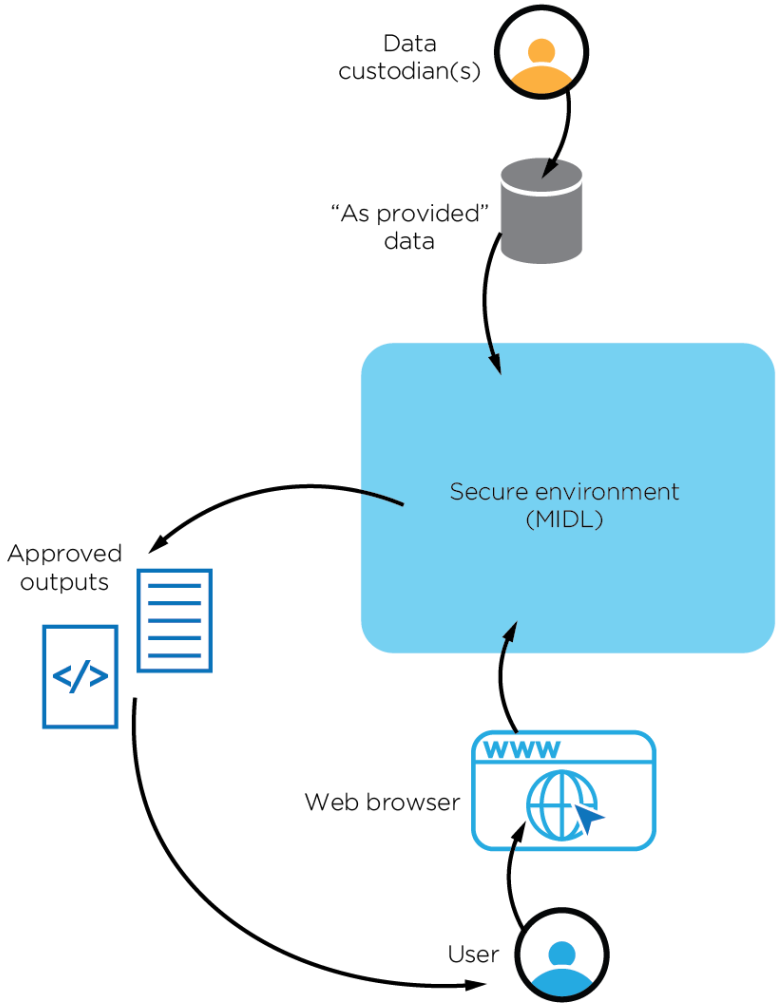
ACSPRI Australian Consortium for Social & Political Research Inc.

What is MIDL?

- MIDL (pronounced ‘middle’) is a purpose-built secure data enclave that enables data curation, analysis and visualization of microlevel data to permit deep and collaborative study of critical issues important to Australian society.
- Established in late 2021 as a collaboration between the Melbourne Institute and Cyconsol, an Australian-based professional services provider.
- MIDL was also supported by a large group of services across the wider University including Business Services, Cybersecurity, Legal & Risk, and Data Governance.
- MIDL was established with funding from the Paul Ramsay Foundation and the University of Melbourne.

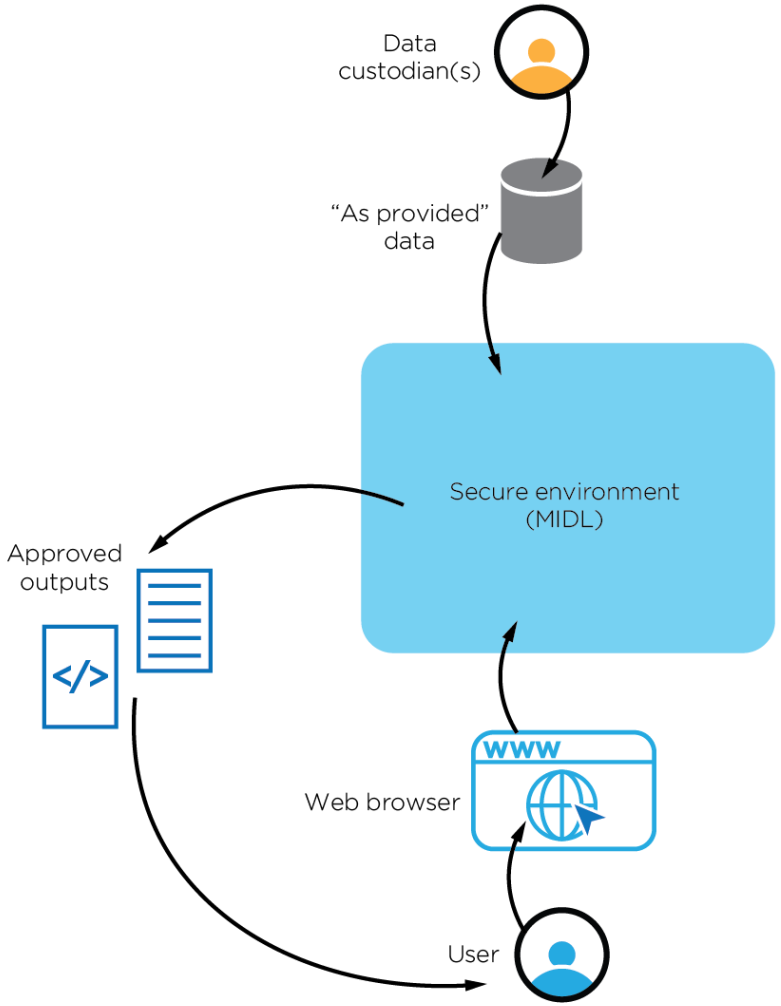
Service offering

- Remote access to virtual environment (Windows) with a range of data analysis and statistical software packages for analysis of sensitive microdata.
- An array of information security controls are implemented as required by Australian Government and international regulations.
  - PROTECTED classification under Aus. Gov. PSPF
- Ongoing assurance activities:
  - IRAP: last assessment in 2022, work underway for 2024
  - Annual penetration testing
  - Broader privacy assessment through PIAs
- Currently undergoing accreditation to be an Accredited Data Service Provider under ONDC DAT Scheme



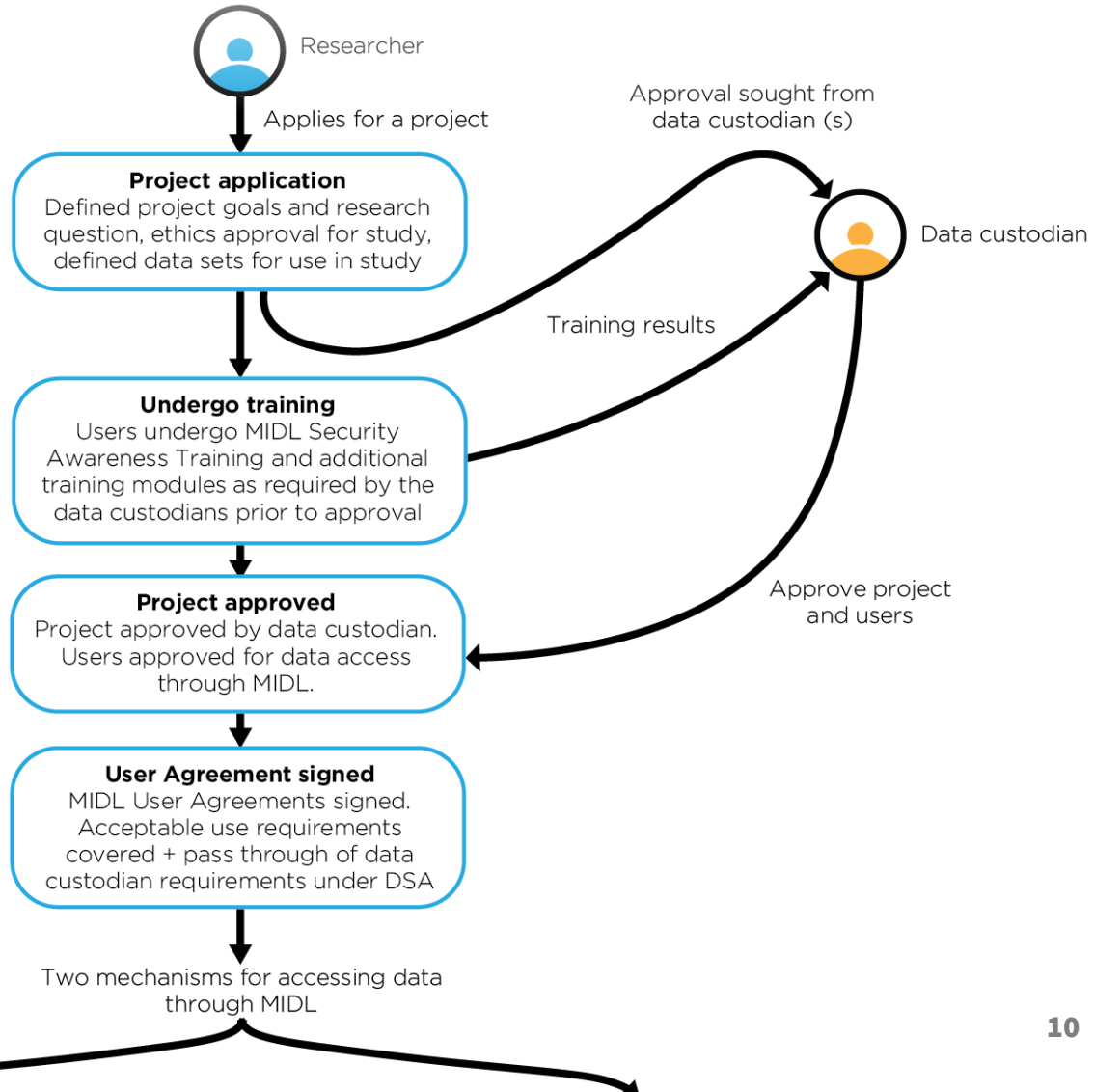
Service offering

- Two modes of access:
  - Projects vs **Shared Data Environments**
- Access to a range of additional resources that speed up research through the setup of Shared Data Environments
  - “Research-ready” data assets that have been carefully curated by the MI’s Foundation Fellow Program with oversight by MI researchers.
  - An in-environment wiki solution (MIDL Wiki) for sharing, documenting project activities.



User journey

Methods of access



User journey

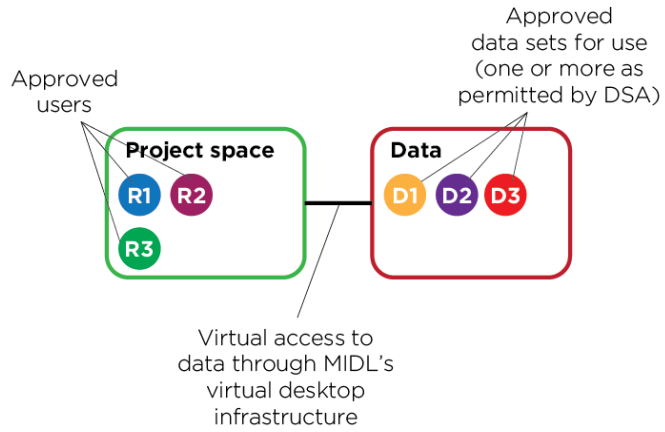
Methods of access

Standalone Project

A standalone project that is defined by a singular research topic (as defined by the user).

A project consists of a workspace accessible by multiple approved users which can access multiple data sets for analysis as per the DSA for those data sets.

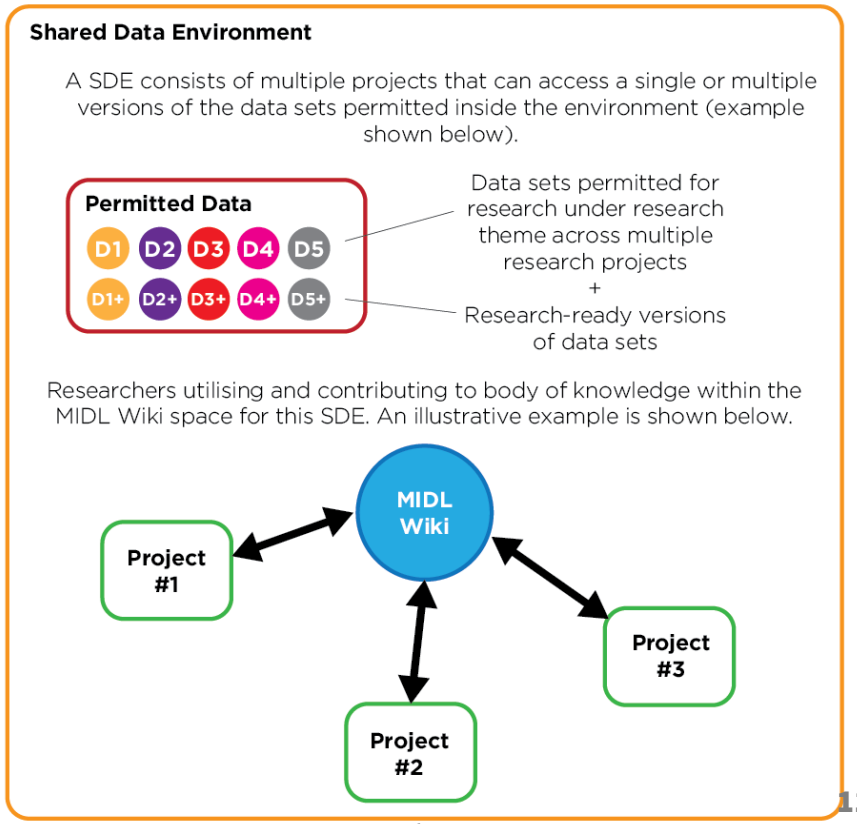
Users can access their data sets within their project workspace using MIDL's virtual desktop infrastructure.



Shared Data Environment

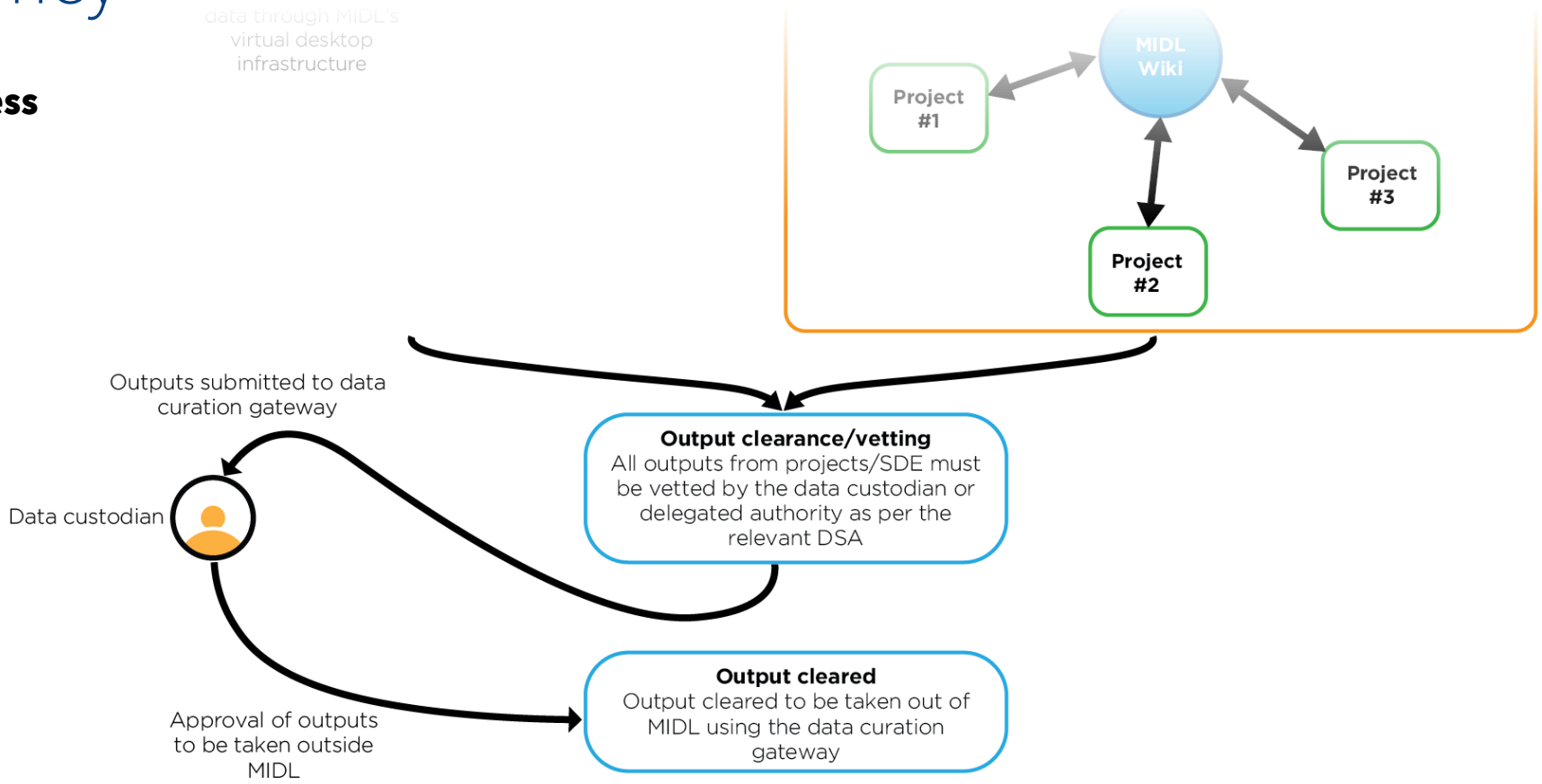
A Shared Data Environment consists of multiple projects with research topics that align with a theme of research.

Projects under a SDE have access to a single or multiple data sets that are selected and curated to be used for evidence-based analysis of questions related to the given theme.



User journey

Methods of access

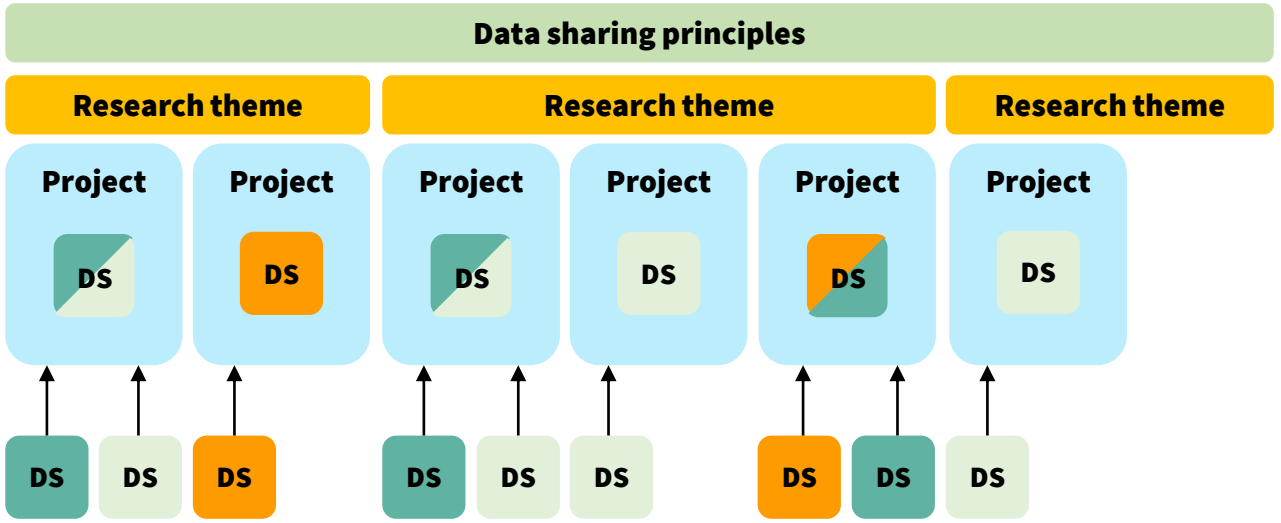


Concept of the Shared Data Environment

Project-based model

Project data* created as part of project activities

"As provided" data from Data Custodians



* Includes "research ready" data, documentation, analysis code, figures, outputs etc.

Concept of the Shared Data Environment

What is a 'research-ready' data set? [1]

"A 'research-ready' data set is a data set that has undergone a range of technical tasks such as data transformation, harmonisation, data cleaning and preparation; as well as standardisation and documentation. The aim of creating such a data set is to do sufficient processing of the data to reduce the technical burden on the researcher and make the data available for **broad research purposes**.

Subsequently, more processing could be performed by experienced researchers in particular sub-themes of research to create an 'analysis-ready' data set that is aimed at **investigating specific research questions**."

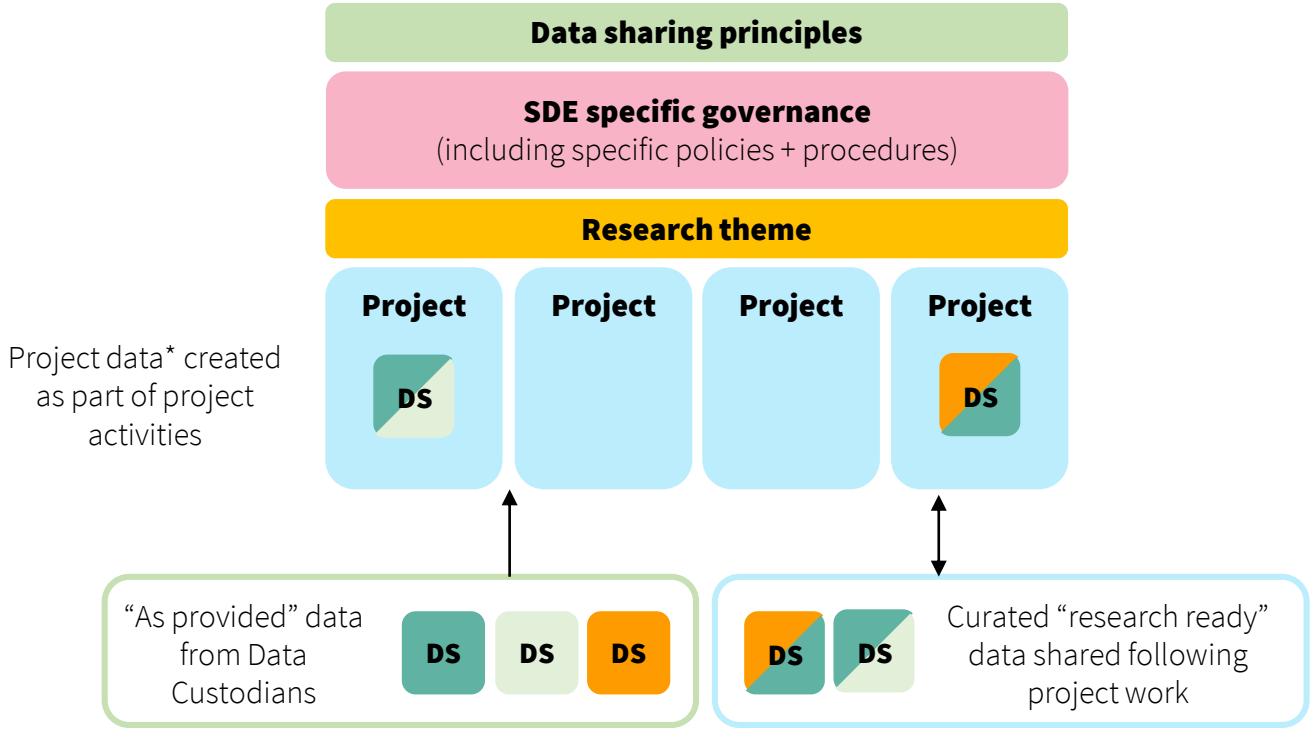
Framework for creating a 'research-ready' dataset[1]

- Formulate the question
- Understand the data
- Process the data
- Documentation

[1]Payne and Samarage, "Maximising evidence-based policy analysis through data sharing" In Dawkins, Peter and Payne, A. Abigail (Eds.) (2022). Melbourne Institute Compendium 2022: Economic & Social Policy: Towards Evidence-Based Policy Solutions. Melbourne Institute: Applied Economic & Social Research, University of Melbourne, Australia

Concept of the Shared Data Environment

The Shared Data Environment – A digital community of practice



Recipe for a “Research Ready” data set memo

Released for data users to access.

Data development (Data custodian/provider)

- Data collection
- Development of data elements
- Data documentation

e.g., ABS undertaking collection and development of the Australian Census, Dept. of Education and NCVER developing the LSAY data, ACARA compiling data on students and schools in Australia.

Not limited to creation of primary data. Applies to administrative data too.

Research ready data development (SDE)

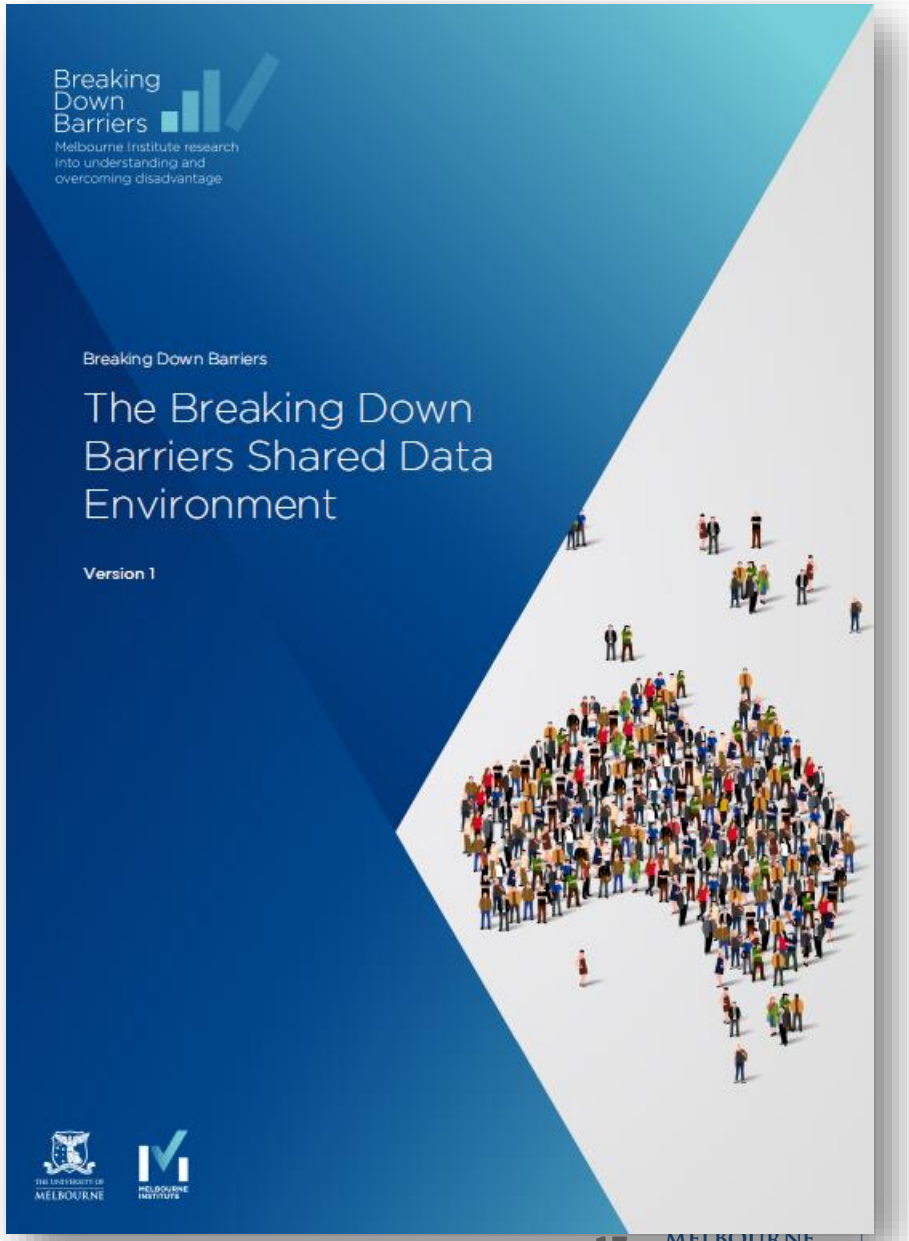
Formulate the research question(s) or themes to support further development

e.g., to support research into poverty and disadvantage BUT more specific than that i.e., to support research into understanding how youth transition into poverty after school.

Check data viability	Does this data add to other data to support the research question? Has a data mapping exercise been undertaken?
Understand data scope	How was the data collected? What is the target population? How would this affect someone using this data to answer the research question?
Understand what the data is about	What measures are captured in the data? Are they relevant? Can they be re-used? Do they need to be modified?
Document decisions	What new variables are defined? What did we change? What decisions were made to back up these changes?
Compile for better user experience	Is information on this data scattered in multiple areas? Compile into a data memo that makes it easier for the user to understand the data.

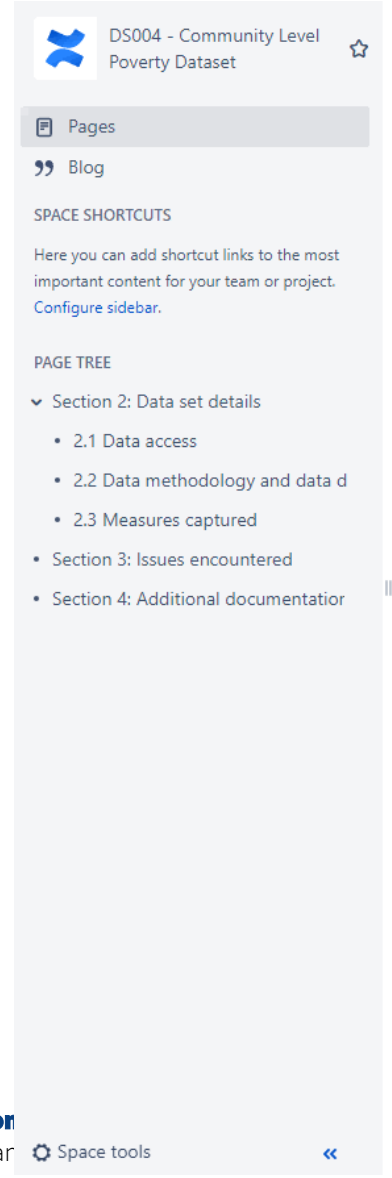
The Breaking Down Barriers SDE

- The Breaking Down Barriers project was the first Shared Data Environment (BDB-SDE) established within the MIDL environment. This large project is funded by the Paul Ramsay Foundation and its aim is to **inform and shape policy and practice to break cycles of poverty in Australia**.
- The BDB-SDE is an environment that brings together data from a range of sources to enable deep analysis, testing and evaluation of ideas to address poverty and disadvantage.
- The BDB-SDE is being built incrementally and focussed on identifying and enabling access to relevant data sets through three mechanisms:
  - Access to a range of broader datasets at Commonwealth and State levels
  - Undertaking short reviews on specific issues tied to disadvantage
  - Supporting initiatives undertaken by service providers and/or other researchers



Concept of the Shared Data Environment

Example memo series from the Breaking Down Barriers SDE



Section 1: Data Snapshot

Key Information	
Brief overview	The Community-Level Poverty Dataset provides information about key demographic, social and economical information at the Statistical Area Level 2 (SA2), Statistical Area Level 3 (SA3) and Statistical Area Level 4 (SA4) level. This dataset was created within the Australian Bureau of Statistics (ABS) DataLab using the Australian Census Longitudinal Datatset (ACLD) 2006-11-16, the ACLD 200611-16 and the 2016 Census Sample File.
Method of access	Currently this data asset is being developed inside the ABS DataLab. When ready, a release version of this data will be provided through the BDB-SDE.
MI contact for key information	Prof. A. Abigail Payne, Dr. Rajeev Samarage
Additional information	Additional information relating to this data asset can be found in the subsquent pages in this wiki space. They can be accessed through the navigation pane on the left hand side of this window.

Detailed Information		
Data set name/abbreviated form: N/A		**Data set name/full form:** Community Level Poverty Data-set
Data custodian name: Melbourne Institute (aggregated form)		**Primary contact name:** See above.
Data set form Aggregated data comprising of person/household/family level statistics	**Unit of observation** Geographical Regions	**Statistical geographical level** SA2, SA3 and SA4
Sampling framework The Australian Census Longitudinal Datatset (ACLD) 06-11-16 consists of a **5% representative sample of persons** from the 2006 Census of Population and Housing. These records are then linked to the 2011 and 2016 Census. However, to create this particular dataset only the data from 2006 is used. The ACLD 11-16 consists of a **5% representative sample of persons** from the 2011 Census of Population and Housing. These records are then linked to the 2016 Census. However, to create this particular dataset only the data from 2011 is used. The 2016 Census Sample File consists of a **5% representative sample of private and dwellings non-private** from the 2016 Census of Population and Housing.		
Years of coverage This dataset is equivalent to repeated cross-sectional analysis. Data is obtained from a different data source for each year, and then		**Data collection method** Primary Data. The Census of Population and Housing is conducted every five years to measure the number of people and dwellings in Australia on Census

Variable Name	Variable Description	Census Questions Used	Census Variables Used	Variable Development
			in the 2016 Census Sample FIle	terms of the 2016 SA4.
sa4_not_inc	Variable description: Some SA4s (very few) are not included in all three Census Years. This variable describes the reasons certain SA4s are not included in all three Census Years. More detail about the values this variable can take is in the Variable Development Section.			One SA4 is not included in all three Census Years. This variable is included in the SA4 dataset only and takes the following values: • Not Applicable (not relevant as this SA4 is included in all three Census Years) • ext_territory means that this is the code for external territories. These are included in the ACLD-06-11-16 and the ACLD-11-16 but not the 2016 Census Sample File.
year	The Census Year the measurement was taken. This can take the values of 2006, 2011 and 2016.	N/A	N/A	N/A

2. Measures Captured

2.1 Population Measures (denoted with a prefix: '**pop_**')

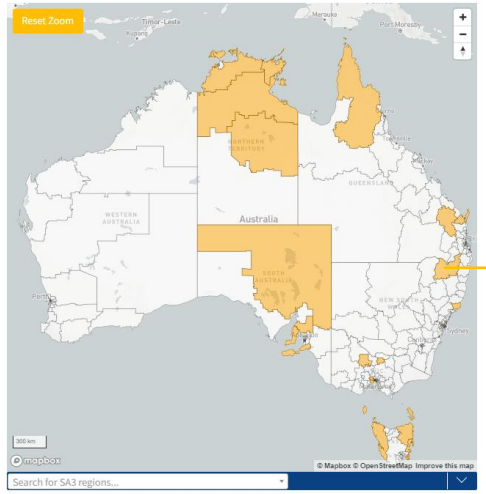
Variable name	Variable description	Census Question	Census Variables Used	Variable Development
pop_tot	Number of people in a SA2/SA3/SA4	N/A	N/A	This variable was created by assigning each unit record in the ACLD-06-11-16, the ACLD-11-16 and the Person 2016 Census Sample File a value of 1
pop_tot_dum	This is a low-population dummy variable.			An SA2/SA3/SA4 is assigned a value of 1 if the total weighted population is less than 500 (unweighted population less than 25) and a value of 0 otherwise
pop_aus_nonindig	Number of people who are born in Australia and Non-Indigenous	• Is <person> of Aboriginal or Torres Strait Islander origin? • In which country was <person> born?	• Country of Birth of Person (bplp_06) and Indigenous Status (ingp_06) in the ACLD-06-11-16 • Country of Birth of Person (bplp_11) and Indigenous Status (ingp_11) in the ACLD-11-16 • Country of Birth of Person (BPLP) and Indigenous Status (INGP) in the 2016 Census Sample File	This variable was coded using the Country of Birth and the Indigenous Status variables from the ACLD-06-11-16 and the ACLD-11-16. Born in Australia includes those who are born in Mainland Australia and Other External Territories. Foreign-born is anyone who stated any other birthplace or described themselves as being born overseas but did not describe the exact country. An Indigenous person is defined as someone who identifies as Aboriginal, a Torres Strait Islander or both an Aboriginal and Torres Strait Islander.
pop_aus_indig	Number of people who are born in Australia and Indigenous			
pop_fb_nonindig	Number of people who are born Overseas and Non-Indigenous			
pop_fb_indig	Number of people who are born Overseas and Indigenous			
pop_ns_inad	Number of people who did not describe their Country of Birth and/or Indigenous Status			

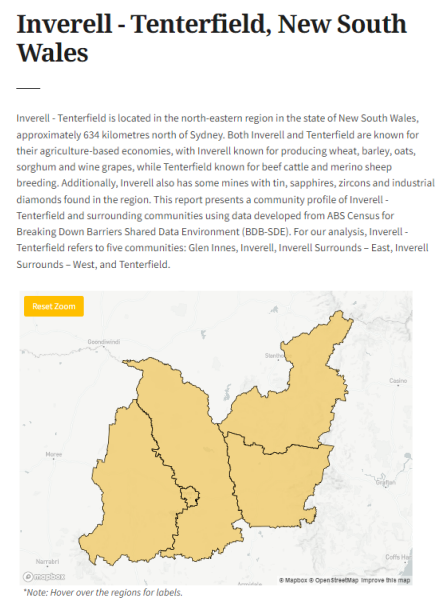
2.2 Household Composition (denoted with a prefix: '**hh** ')

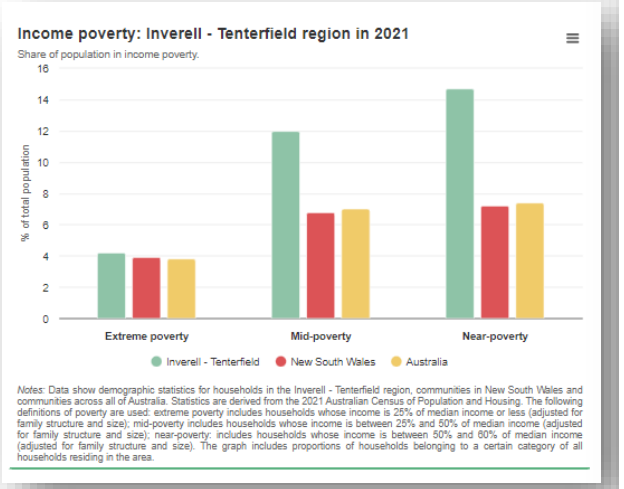
Recent works from SDEs

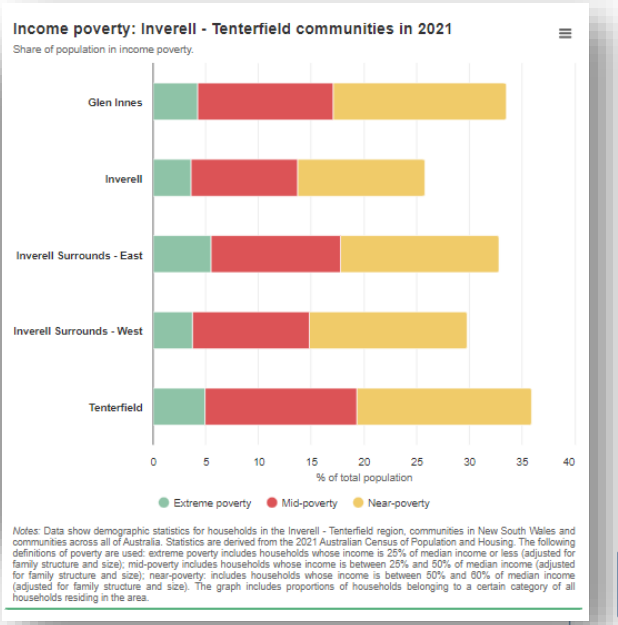
Breaking Down Barriers Community Profiles

- An interactive web-based application that provides socio-economic data and key insights into regions and communities across Australia.
- Preliminary data concept > BDB demonstration report that investigated community-level poverty (Payne and Samarage, 2020)
- Careful curation of data and measures from a range of data sources as part of continuous updates to profiles:
  - Australian Census data: 2006 to 2021
  - Service provider data: 2013 to 2021
  - Youth employment index (currently being developed by MI)
  - School-level and institution-level data on education
  - And more…
- Phase 1 live now:
  - https://bdbprofiles.melbourneinstitute.unimelb.edu.au/



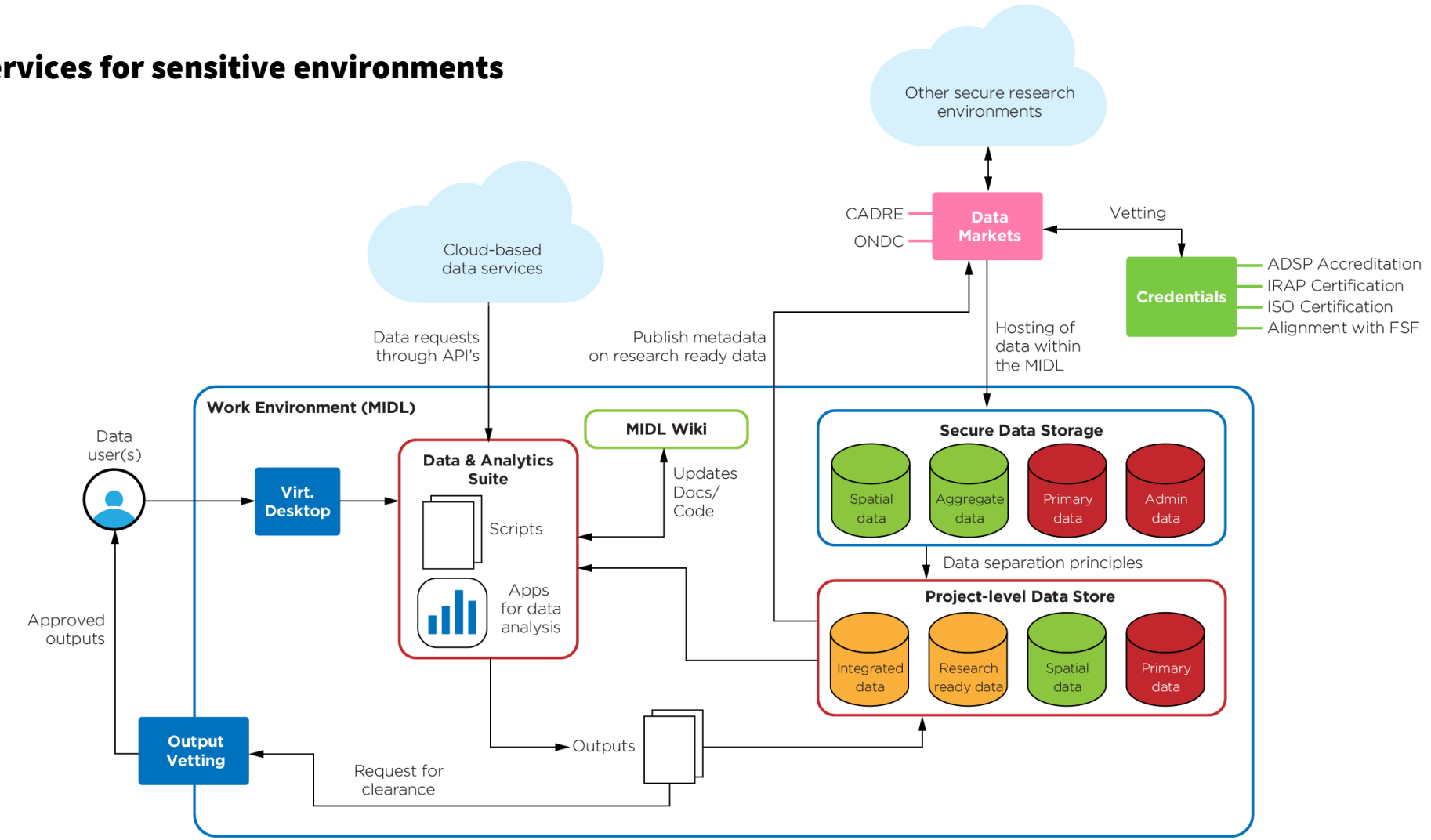






Phase 2 and beyond

Integrated services for sensitive environments



Closing remarks

- There is a growing need to cater for granting access, and enabling better use/re-use of data from both primary and secondary sources.
- There is no “one size fits all” solution when it comes to data access environments and data services available for research and analyses.
- In the end, enabling faster and ‘richer’ research using a broad range of datasets with minimal risks to data owners, is critical.

Sensitive data integration services using shared data environments

Dr. Rajeev Samarage

Melbourne Institute: Applied Economic & Social Research, The Faculty of Business and Economics
The University of Melbourne

eResearch Australasia 2023 Conference
16 – 20 October 2023, Brisbane, Australia

THE UNIVERSITY OF MELBOURNE

MELBOURNE INSTITUTE
APPLIED ECONOMIC & SOCIAL RESEARCH