Social Sciences Co-design Workshop Report for the HASS and Indigenous Research Data Commons

The Australian Research Data Commons

4/7/2024

DOI: 10.5281/zenodo.12639457

Australian Research Data Commons

CONTENTS

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative
ARDC is enabled by NCRIS

EXECUTIVE SUMMARY

From 2022, the ARDC has been making a major investment in digital research infrastructure for humanities, arts, social sciences and Indigenous research data communities. This work has been done in accordance with the 2016 NCRIS Roadmap and Department of Education commissioned studies. The projects undertaken in this area formed an emerging *HASS and Indigenous Research Data Commons*.

In late 2023, the ARDC received the largest ever investment in HASS and Indigenous Research Capability. In Early 2024 we initiated a process of co-design to deliver on an expanded HASS and Indigenous Research Data Commons, covering existing and new areas of research infrastructure.

Following the completion of the Integrated Research Infrastructure for Social Sciences project, ARDC is looking to develop new digital research infrastructure for the Social Sciences. This infrastructure development will build upon a successful pilot project that is improving valuable administrative data for research use, and is informed by the direction set by the Academy of Social Sciences in Australia's Decadal Plan for Social Science Research Infrastructure 2024-33.

The ARDC is co-designing its new infrastructure investments with the research community and other stakeholders to ensure that we create the greatest possible impact. Development of this infrastructure follows the HASS and Indigenous RDC co-design framework.

The ARDC conducted two co-design workshops in April 2024 to shape plans for this investment opportunity. These workshops were open to all, and were attended by a total of 83 participants from 38 organisations, including Australian and international universities, Federal Government departments, research institutes, peak bodies, research infrastructure providers, and GLAM sector organisations. In Workshop 1 we refined our understanding of the challenge to be addressed and found out what outcomes participants hoped to see. In Workshop 2 we explored how the desired outcomes could be achieved within the new infrastructure investment.

THE HASS AND INDIGENOUS RESEARCH DATA COMMONS

The ARDC is planning major investment in digital research infrastructure for humanities, arts, social sciences and Indigenous research data communities.

The 2016 National Research Infrastructure Roadmap identified avenues to enhance the impact of HASS (Humanities, Arts, and Social Sciences) and Indigenous research, proposing improvements in coordinating research infrastructure to facilitate access to and analysis of physical and digital collections. These improvements involve leveraging tools like digitization, aggregation, and interpretation platforms.

Subsequently, the Australian Government Department of Education commissioned three studies to pinpoint investment-ready programs eligible for National Research Infrastructure funding. While not all recommendations from these studies received funding initially, the identified activities demonstrated a high level of readiness to engage with and benefit from the development of a HASS and Indigenous Research Data Commons (RDC).

From 2022-2024 the ARDC-led HASS and Indigenous RDC has implemented a number of these activities, including:

- Improving Indigenous Research Capabilities (IIRC) led by Dist Prof Marcia Langton, University of Melbourne
- Establishing the Language Data Commons of Australia (LDaCA) led by Prof Michael Haugh, University of Queensland
- Developing Integrated Research Infrastructure for Social Sciences (IRISS) led by A/Prof Steven McEachern, Australian National University
- Creating the Trove Researcher Platform, which comprised Trove Enhancements (led by ANU) and the ARDC Community Data Lab (led by the ARDC)

In 2023, the ARDC received its largest-ever investment in HASS research infrastructure and Indigenous research capability. This $25 million grant, part of the Australian Government's 2023 National Collaborative Research Infrastructure Strategy (NCRIS) Research Infrastructure Investment Plan Funding Round, supplemented by co-investment from national partners, will further establish long-term, national digital research infrastructure to support HASS and Indigenous research data communities in Australia.

Our commitment extends to supporting existing activities such as IIRC, LDaCA, the Community Data Lab, and Social Sciences while also considering the expansion of the RDC to include new activities for Creative Arts and Media(ted) Data. It has become evident that sustained investment is crucial to scale up these activities and provide coordinated digital research infrastructure meeting the needs of humanities, arts, social sciences, and Indigenous researchers.

Phase 2 aims to consolidate previous efforts, capitalise on synergies between existing and new projects, broaden partnerships and disciplinary coverage, and deepen engagement with the GLAM sector and Indigenous data governance.

In developing new activities, we have considered researcher needs gathered from consultations in 2021 and identified areas with active, coherent communities where research infrastructure can flourish. We have also looked at existing infrastructure that could be bolstered to transition into enduring national assets.

THE SOCIAL SCIENCES INVESTMENT OPPORTUNITY

Following the completion of the IRISS project in early 2024, the ARDC has been investigating new opportunities for infrastructure investment in the social sciences.

In March of 2024, the ARDC began a pilot project to investigate the possibility of working to improve the access and usability of government administrative data for social sciences research. Data generated by government agencies and service providers during their daily operations can be a rich source of information about many aspects of Australian society. However, because this data is not generated with the primary intention of use in research, it can take considerable time and effort to make it research-ready. In this pilot project ARDC is working with the Institute of Social Sciences Research (ISSR) at the University of Queensland and the Australian Bureau of Statistics (ABS) to improve the metadata of Higher Education administrative data held within the ABS Person Level Individual Data Asset (PLIDA). These metadata improvements will make the data more usable for social science researchers. This pilot (which will be completed in August of 2024) has demonstrated the success of the ARDC-ISSR-ABS partnership, and there is potential for substantial further work to improve administrative data for research use.

The ARDC's planned infrastructure investment for the social sciences will also be guided by the Decadal Plan for Social Science Research Infrastructure 2024-33 recently released by the Academy of Social Sciences in Australia (ASSA). This plan outlines a vision for transforming Australia's social science research infrastructure over the coming decade.

The proposed plan for the Social Sciences is being developed in collaboration with the ISSR at the University of Queensland. Collaborating organisations include the ABS, the Australian Department of Education, Monash University, Australian National University and the University of Western Australia.

OUR CO-DESIGN PROCESS

The ARDC is developing this infrastructure in accordance with the HASS and Indigenous RDC co-design framework (https://doi.org/10.5281/zenodo.10516606). This process is designed to create the greatest impact for research and researchers by co-designing the infrastructure with the people who will benefit from it. Development will follow the stages of Problem Identification, Project Shaping, Project Planning, and Endorsement.

Problem Identification has taken place through extensive consultations and information gathering sessions, as described above. By tracking emerging gaps, opportunities, and sectoral trends, the ARDC has identified potential avenues for delivering benefits to the HASS and Indigenous research community.

This strategic delineation of opportunities reflects the ARDC's commitment to addressing pertinent challenges and fostering innovation within the HASS and Indigenous research landscape.

The Project Shaping phase was implemented through the two open co-design workshops described below. Workshops were advertised through ARDC newsletters, social media channels and other networks. We sought to involve experts in research practice, disciplinary needs and infrastructure provision to better understand the current challenges faced by stakeholders and the outcomes that they hope to see, and to shape potential solutions that could be provided by the new infrastructure.

Project Planning and Endorsement phases will follow - further information is provided in the "Next Steps" section below.

WORKSHOP 1

The first co-design workshop was held alongside the launch of the ASSA Decadal Plan for Social Science Research Infrastructure 2024-33 on April 10th 2024. This was a hybrid event with attendees both in person in Canberra and online. It was attended by 51 participants from 26 organisations. Organisations represented included 11 Australian universities, as well as Federal Government departments, research institutes, peak bodies, research infrastructure providers, and GLAM sector organisations.

The aim of this first workshop was to understand the challenges being faced by social sciences researchers and identify high-priority challenges that could be addressed through this investment.

Tomasz Zajac from the ISSR team set the scene for the workshop by describing recent exemplar digital research infrastructure projects for the social sciences, including the GeoSocial Data Integration Service delivered as part of the IRISS project, and the Enhancing Metadata for Inclusive Research on Entrenched Disadvantage pilot project. Mark Western then spoke on infrastructure development opportunities identified in the ASSA Decadal Plan.

In the first activity, participants formed breakout groups and brainstormed answers to the following questions, recording their responses using the Miro online whiteboard tool:

1. What are the most pressing data/digital infrastructure challenges for social sciences research?
2. What could we achieve in the social sciences with data and digital infrastructure? What are the opportunities?
3. What challenges/opportunities could your organisation contribute to tackling?
4. What work is already taking place/what already exists that we can build on?

Together, the participants then reviewed all of the answers on the board and arranged them so that similar ideas were clustered together. In their breakout groups they discussed the clusters that had emerged, and each group identified a shortlist of promising opportunities for future work.

To conclude the session, participants used mentimeter to individually report back on the potential project ideas from the discussion that excited them most.

Based on the responses from both mentimeter and the group discussions, participants had the following high-level objectives for an infrastructure investment for the social sciences:

- Make use of existing data assets
    - Improve discovery of and access to existing data
    - Improve data and metadata quality
    - Link existing data assets
    - Increase interoperability and uptake of existing research infrastructure
- Support a skilled workforce
    - Uplift technical skills of social sciences researchers
    - Create a pool of skilled software engineers with social sciences expertise
    - Provide career progression and stability for technical workers in the social sciences
- Forward-thinking infrastructure development
    - Consider the resources required for data-related work
    - Plan for maintenance and sustainability of infrastructure
- Advocacy and public licence
    - Aim to raise the profile of social science expertise
    - Increase public licence by creating public-facing applications of the data/infrastructure

The workshop also surfaced a number of more specific development opportunities:

- Lifecourse data trajectories: Cradle-to-grave data trajectories with ability to drill into community-level data, combining data from government, NGO and research
- Development of a national longitudinal data asset covering adults 65+
- Tools and data to evaluate impact of policy
- Integration of Household, Income and Labour Dynamics in Australia (HILDA) survey data into the Person Level Integrated Data Asset (PLIDA).

Full responses can be found in Appendix 1.

A GRAND CHALLENGE FOR THE SOCIAL SCIENCES

During discussion of the ASSA Decadal Plan in the roundtable events held alongside Workshop 1, participants had raised the idea of identifying grand challenges that provide a common direction and demonstrate impact of work in the social sciences. To help focus our discussion going forward, ARDC and ISSR developed the following grand challenge statement:

Enhancing data infrastructure for a more equitable and resilient society

Australia faces a national challenge to build a more equitable and resilient society, communities, and regions. The key question is: what new, large-scale research infrastructure, including data assets, analytic tools, and research infrastructure services will be required to support researchers to address this challenge?

SURVEY

We sent a survey to all participants of Workshop 1 to introduce the idea of the grand challenge and collect further input.

We received 24 responses, including respondents from 12 Australian universities, one international university, and two research institutes.

In the survey, we asked respondents to consider how their work in the social sciences contributes to addressing the grand challenge of building a more equitable and resilient society, and what data or digital research infrastructure they currently use in that work.

Respondents reported using the following data sources:

- Administrative data (including PLIDA, DEX, BLADE)
- Longitudinal study data (including HILDA, LSAC, LSIC, LSAY, QILT, BNLA, cohort studies)
- Spatial data sources (unspecified)
- Various other sources including NLA, NAA, Trove, Parliament data, Health sector electronic record data, Wikimedia and Wikidata

Other digital research infrastructure reported included:

- Cloud computing services
- HPC
- Coding and statistical languages and packages
- Advanced machine learning

Full responses can be found in Appendix 2.

WORKSHOP 2

The second co-design workshop was held virtually on April 23rd 2024. It was attended by 46 participants from 27 organisations. Organisations represented included 15 Australian universities and an international university, as well as Federal Government departments, research institutes, peak bodies, and research infrastructure providers.

Based on the responses from Workshop 1 and the survey, the ARDC and the ISSR identified a set of potential directions for the Social Sciences infrastructure development. These were:

1. Improving data discoverability (for instance, by developing a "one-stop-shop" for initial information on different types of data available for social sciences research)
2. Enhancing the accessibility/usability of administrative data (by improving the documentation and metadata for key administrative data assets such as the ABS's PLIDA)
3. Further integration of (individual-level) data on people (for instance, by integrating major longitudinal studies with PLIDA)
4. Cross-unit data integration (bringing together data across different units, such as individual, family/household, community, etc, including spatial data)
5. Connecting data over the lifecourse (for instance by connecting existing data sources about different groups of individuals to give a picture of life course trajectories)
6. Considering Indigenous data governance in the context of social science data (working out how to incorporate Indigenous data governance principles into integrated public sector administrative data)

Participants completed a series of discussion activities in breakout groups, and recorded their input using the Miro online whiteboard tool. Some key themes from the responses are highlighted below. Where responses related to a particular development direction identified above, this is indicated. Full responses can be found in Appendix 3.

Activity 1: Benefits

What disciplines could make use of this?

- Improve discoverability
    - General increase in disciplines that can access and use
- Enhance accessibility, Individual-level integration, Cross-unit integration, Life course
    - Sociology
    - Geography

    - Urban planning
    - Environmental science
    - Criminology
    - Demography
    - Political Science,
    - Economics
- Individual-level integration, Cross-unit integration, Life course
    - Cross-disciplinary, cross-sector, cross-domain
    - Social science and health, health policy
    - Social science and natural/built environment, architecture, urban planning
- Cross-unit integration
    - Policy impact assessment
    - Spatially integrated social science
- Lifecourse
    - Adult education

What could be enabled by infrastructure development of this type? What research could be made better? What new questions could be asked?

- Individual-level integration
    - Study trajectories through school, into post-secondary ed, labour market
    - Role of schooling on opportunities for reducing poverty
    - Better control of family background, SES when studying range of life outcomes (health, labour market, family formation, etc.)
    - Factors correlated with low-performing primary students who succeed later in life
- Life course
    - Timelines for maltreatment impacts (not finishing school, homelessness, removal of children)
    - Understand importance of different types of potential interventions for addressing poverty
    - Impact of environmental degradation on in-utero foetus health and later life outcomes

    - Intergenerational trauma
    - Effectiveness of interventions
- Indigenous data governance
    - Better understanding place of Indigenous people in Australian society
- Individual-level integration, Cross-unit integration, Life course
    - Research areas raised repeatedly across responses: Spatial, cross-cultural, family/intergenerational, education, income and poverty, health, migration, housing.

Other benefits identified: time and efficiency

- Improve discoverability
    - Save researcher time in discovery, esp. Students
    - Strong support for a “one-stop shop” to give visibility over multiple data sources, databases (mentioned in 7 responses)
- Enhance accessibility
    - Documentation - decide if data suitable before requesting access
    - Shared experience - learn from challenges and solutions of other analysts on commonly-used datasets
    - Shared code libraries and analysis handbooks - avoid duplication of effort

Other benefits identified: Streamlined access to multiple data sources

- Enhance accessibility, Individual-level integration
    - Streamline/lower barriers of access to multiple data sources - access when researchers need it

Other benefits identified: Improving experience of new users

- Improve discoverability
    - Help students and ECRs find data - particularly those from international backgrounds
    - Seed new projects
    - More mixed methods projects from qualitative researchers
- Enhance accessibility
    - Improved documentation and metadata for some PLIDA modules would increase accessibility for unfamiliar users, esp. from different disciplines

Other benefits identified: Value of integrated data

- Individual-level integration, Cross-unit integration, Life course
    - Strong support - integration of datasets is extremely valuable, allows new research questions (mentioned in 10 responses)
    - Endorsement of importance of both spatial integration and life course view of data

Other benefits identified: Extracting new information from existing data

- Improve discoverability
    - Use AI to capture key concepts and features in unstructured data
- Individual-level integration
    - Identification of population subgroups easier with integrated data
- Lifecourse
    - Support more robust study of factors with very long-lasting impacts

Other benefits identified: Identifying and filling gaps

- Life course
    - Map out what is available over each life stage to identify gaps to be filled (links to Improving Discoverability)
    - Gaps should be filled - e.g. remote areas for HILDA, census data for housing and community planning

Other benefits identified: Ask-first protocols

- Indigenous data governance
    - Database of Ask First protocols
    - Centralised repository that allows Ask First protocols

Other benefits identified: miscellaneous

- Improve discoverability
    - Existing data assets more FAIR, better used
- Enhance accessibility
    - Can be used to make informed decisions about government policies and programs, R&D
- Indigenous research governance

- Better understanding of indigenous data governance
- Community involvement in research
- Enhanced social licence for Indigenous research
- Support implementation of CARE across HASS&I RDC
- Contribute to modern, participatory social science
- Support Indigenous-focused research using administrative data

Considerations raised: Privacy and legal requirements

- Individual-level integration
  - Consider privacy and ethical implications of linkage and integration
- Enhance accessibility
  - Access considered within context and privacy rules

Considerations raised: Related activities in Indigenous data governance

- Indigenous data governance
  - Australian Government Data Catalogue
  - Indigenous Data Network
  - LDaCA

Considerations raised: Custodianship/governance

- Individual-level integration, Cross-unit integration
  - Requires strong relationships with data owners, custodians
  - Integration will create custodianship questions
  - Consider finding a trusted org to govern use
- Indigenous data governance
  - Indigenous ownership, control and access of data

Activity 2: Requirements and Enhancements

What would you require for this to be useful to you/the people you work with? Are there related activities that would enhance this potential infrastructure development?

Metadata

- Improve discoverability, Enhance accessibility, Individual-level integration, Cross-unit integration

- Strong support for importance of metadata (mentioned in 17 responses)
- Standardised metadata across datasets, domains
- Metadata fields requested
  - For datasets: Coverage (spatial, temporal), unit of observation, update frequency, automatic or manual, tags, themes, organisation, how to apply for access, cost, past usages, domain, stage of life course, type of data (administrative, survey), population, linkage with other datasets,
  - For columns: Coverage (spatial, temporal), type, description, measurement unit, foreign key
- Separate metadata from data access
  - For sensitive data - make as much metadata as possible viewable before granting access
  - PLIDA - allow metadata and documentation to be viewed outside of DataLab, without requiring training and approved access
- Present information about data in a way that is understandable across domains
- Automatic validation of metadata. Tools: DBT, Great Expectations

Search

- Improve discoverability
  - Powerful search engine (ElasticSearch, Solr)
  - AI assistance with search over different terminology
  - Search filters requested: Domain, outcome types, stage of life course, type of data (administrative, survey), access conditions (can download now), data compatible with a particular analysis workflow

Map data sources

- Improve discoverability
  - Map who is making what data discoverable, including overlaps
  - Map sources of data which measure the same concepts in different ways

Data quality and limitations

- Improve discoverability, Enhance accessibility

    - Information about data quality - known issues and pitfalls, adequacy for particular kinds of research
    - Information about data limitations, especially coverage
    - Forum where users can freely discuss issues about the data

Prior/potential usage of data

- Improve discoverability, Enhance accessibility
    - Audit what data sources are being used by social scientists in Australia and for what purposes
    - Indication of potential applications of data
    - Examples of previous research/index research outputs for particular datasets to demonstrate previous uses, avoid duplicated effort around methods

Classifications, definitions and vocabularies

- Improve discoverability, Enhance accessibility, Individual-level integration
    - Standardise classifications and definitions
    - Document inconsistencies in classifications and definitions and reasons for them, eg. definition shifts in service provision over time, administrative business rules, different measurement of the same concept
- Indigenous data governance
    - creation of community-driven standards and classifications. for instance supplement ASGS with Indigenous community and language boundaries

Documentation of administrative processes behind data

- Enhance accessibility
    - Improved description of administrative processes behind data, how data is collected

Data standards

- Enhance accessibility, Cross-unit integration, Life course
    - Agreed standards required for collation of data
    - Require organisations to collect data in a usable form
    - Automatic validation of data. Tools: DBT, Great Expectations

Reduce duplication of effort

- Enhance accessibility, Individual-level integration, Life course
    - Index research outputs from datasets, share relevant publications including technical papers - learn from previous methods
    - Centrally derive commonly required variables and hold within or alongside source data
    - Scripts to clean standard/commonly accessed datasets
    - Mechanisms to share code between researchers within DataLab
    - Code libraries for data management, variable construction, measure development

Data custodians

- Enhance accessibility, Individual-level integration, Cross-unit integration
    - Strong relationships with data custodians vital to build trust, enable access, encourage future developments
    - Need to consider custodians' needs for security

Permissions and access management

- Improve discoverability, Enhance accessibility, Cross-unit integration
    - Need for a central access management point
    - Ability to conduct linkages with pre-determined permissions from custodians
    - CRUD permissioning in an hierarchy within central team and data custodians
- Indigenous data governance
    - Contextual data access

Access environments

- Enhance accessibility, Further integration, Life course
    - Ability to access administrative datasets in shared environments with access to standardised tools
    - Ability to access data outside of a single/few secure environments
    - Capacity to accredit appropriate, external analysis environments to effectively combine administrative data with research data rather than pushing data between environments
    - Capacity to manage big and complex data

Tools

- Improve discoverability, enhance accessibility

    - Standardised tools embedded within access system
- Individual-level integration, Cross-unit
    - Sustainable access to well-maintained tools
    - Tools that offer data curation at range of spatial and temporal scales
    - Interactive dashboards/tools to map spatial data overlaid onto survey data
    - Tools to enable efficient, robust, repeatable linkage

Training

- Strong support for training overall - mentioned in 13 responses
- Enhance accessibility
    - Training in: general quantitative skills for social sciences researchers, safe/ethical use of data, working with analytic methods in DataLab, use of existing standard data assets.
- Individual-level integration
    - Training in use of longitudinal data, particularly when integrated with administrative data
- Cross-unit, Life course
    - Training to support cross domain research, introduce researchers in one field to methods, data approaches in another
- Indigenous data governance
    - Training for non-Indigenous researchers in Indigenous data governance and research practices and principles
    - Culturally-informed research training for Indigenous researchers
    - Build capacity for data governance with elders and community

Desired linkages

- Improve discoverability
    - Comprehensive coverage in terms of topics/domains
    - Measures of: housing, finance, environment, disability
    - Data with breakdown by city
- Individual-level integration
    - Measures of: education (full histories), assets, money, education, employment, housing, accessibility to healthy options

- Datasets cross-referencing LGAs with categories of regional and rural areas
- Integration of existing cohort data

- Cross-unit integration
  - Integrate QILT surveys (SES, GOS) with TCSI/PLIDA
  - Family structures (to link individuals within family) or link individual data to family characteristics
  - Ability to use multiple Census rounds with a single dataset in PLIDA (capture family characteristics at different points in life)
- Life course
  - Spatially-informed data about inequality, mobility, regional economic development, responses to climate change, community resilience
  - Longitudinal study of ageing
  - Full educational histories
  - Matching late life adult education populations with early school leavers, broken education participation
  - Places, land uses, assets such as private vehicle fleets

Establish Australia-specific Indigenous data governance protocols

- Indigenous data governance
  - Australian version of Canada's OPAC
  - Existing resources:
    - https://www.maiamnayriwingara.org/
    - https://www.lowitja.org.au/news/taking-control-of-our-data-a-discussion-paper-on-indigenous-data-governance-for-aboriginal-and-torres-strait-islander-people-and-communities/

Other requirements and observations

- Improve discoverability
  - Consider contributing to Australian Government Data Catalogue
  - AI techniques for discovering concepts and features not traditionally recorded in structured data

- Enhance accessibility
    - Governance standards, agreements
    - Expert support - facilitate discovery of experts with experience working with these datasets for research purposes
    - A national version of Victorian Social Investment Model (VicSIM)
- Individual-level integration
    - NIHSI AA data as a model for integration of state and commonwealth education data
- Cross-unit integration
    - Support for exploratory discussion across disciplines
    - Database of qual studies using quant datasets
    - Use of data hubs
    - Protocols to bring in other datasets
    - Three data environments - development, staging, production
- Life course
    - There are many ways to connect - e.g. person, household, cohort, group, community, etc. level. Not all data sets will be linkable at an individual level. Connecting data over the life course will involve different ways to link information across data set
    - Much focus is on government data. For long-term success, mechanisms are needed to enable linking of different data sets, of verification of the data, as well as to enable creation of new data (field experiments, surveys, etc.) that can be added

Activity 3: Connections and Gaps

What are the relationships and overlaps between the possibilities we've discussed today? Are there any opportunities we haven't discussed so far in today's session?

Working with data custodians

- Heavily-discussed theme - mentioned in 17 responses
- Trust between researchers and data custodians
    - Considering requirements to build that trust (e.g. DATA Scheme accreditation)
    - Clarity for custodians about how data will be used (technology to provide visibility for custodians over usage in different projects?)

- Connections and trust between custodians (required for linked assets)
    - Seems to be missing - could a third-party data linker help?
- Value proposition for custodians - how do we return products of research to benefit them?
- Sharing experiences of requesting access and what was required from a given custodian to speed process
- Identify mid- and low-level data analysts in large organisations/departments - they understand idiosyncrasies and opportunities for aggregation
- Can be difficult to access Commonwealth and data data if not working on a government project.

New technologies

- Localising LLMs for local use
- Ecosystems of different kinds of data - combine traditional and new (AI) techniques
- Next horizon - new sources, representative synthetic data

Working across disciplines

- Encourage cross-disciplinary work
- National appraisal of Big Questions in various HASS fields to identify who holds data that could be useful to others

Linkage of new data sources

- From other ARDC focus areas (such as social media data from the Australian Internet Observatory)
- Linkage of social science and non-social science datasets
- Beyond government data to industry, non-profits

Data quality

- Quality of data varies between sources
- Establish one dataset that sets the gold standard for metadata, access, quality, and bring others to that standard
- Automatic quality validation

Technical underpinnings

- Secure data access
- Suitable software/computing environments

- Versioning system for data, metadata
- Pipeline management system for automatic update of high-frequency data (AirFlow, Prefect)

Privacy and ethics

- Linking cohort and admin data - ethics and governance requirements
- Risk of public backlash about identification and privacy around linkage

Community and collaboration

- Need to build cross-domain relationships
- Community of practice around use of administrative data
- Working more effectively together to combine efforts
- https://scccp.net/resources/cobeo-tools/

Potential partners to involve

- AIHW
- ONDC

NEXT STEPS

The ARDC has worked with the potential project team to develop a draft project plan, released for public feedback alongside the publication of this report. Feedback submissions will close at the end of day, Thursday 18th July. To submit feedback on project plans please use the form on the ARDC website. All feedback will be collated and responded to in a published report, alongside an updated project plan taking into account this feedback wherever possible. Work on this project is planned to commence in October 2024.

The development of the resulting infrastructure will involve ongoing public testing and consultation - please make sure you are signed up to the HASS and I newsletter so that you receive notifications about upcoming opportunities to be involved.

If you have any other comments or questions about this process please contact us at hassi.codesign@ardc.edu.au with the subject line “Social Sciences”.

APPENDIX 1: WORKSHOP 1 RESPONSES

Activity 1: Brainstorming opportunities

QUESTIONS	RESPONSES
What are the most pressing data/digital infrastructure challenges for social sciences research? What could we achieve in the social sciences with data and digital infrastructure? What are the opportunities? What challenges/ opportunities could your organisation contribute to tackling? What work is already taking place/what already exists that we can build on?	• Data impacting Indigenous peoples easily accessible in one place • Long-term storage/access • Standardized data collection, storage, and access for research purposes for both existing and new data sources (e.g., non-profits admin data) • Data access, repositories, storage • Data access and data standards. • Timely, relevant training materials on Data Ethics & Indigenous data principles for researchers • Research ethics committees not being a barrier to innovating with AI responsibly • intern data engineers and train them up on integration, assembly, pipeline development. We'll definitely get a benefit in return! New ideas state of the art techniques and approaches etc. Could be a pragmatic collaboration. • Analytical tool dev and training • Linked administrative data from cradle to grave • assessing value of what data to link, keep or discard • sensitive data governance & sharing platforms that can be repurposed for Indigenous data governance • Stop the loss of public investment in research data by providing proper curation so data is not lost once projects end. • Commercial ownership of important data • PUBLIC FACING RESEARCH MATERIALS • New questions opened up by comparing data across domains and jurisdictions • size and storage • Licensing of curated data • privacy concerns • Longevity of curated data • International comparisons

QUESTIONS	RESPONSES
	• novel research questions and methods • better ROIs for investment • connected multi-org approach to infrastructure funding strategy across NCRIS, LIEF and other funders • Quality qualitatitve data • integrating quant and qual data in evaluation • Vendors own more social data than we do.. Digital Observatory challenge • insufficient metadata • Consistent metadata specification • Understanding of longitudinal dynamics of human development pathways • Technical staff to build and then maintain RIs (+1) • Lack of resources for making research software user-friendly and maintaining it • Capability Availability of Data and Platforms • Slow down the research cycle, include support for latest tech, with aims to support quality, continuity and reusability • social sciences research directions and methodologies are heterogenous • Visibility which promote true cross cutting capability • Understanding important (yet relatively small in number) sub-populations and social problems • Investigating Indigenous Data - best practice principles - continue to expand on this. Training. Ensure the university understands the importance • The UVP for social science within data ecosystems and data platforms • Do we start sector by sector (Universities, Whole of Gov, State) Indsutry • cost • time • New areas of enquiry can be explored • managing the volume of data - tools, discovery eg Decadal Plan - auto-summarise (AI??) • Students on a living wage

QUESTIONS	RESPONSES
	• accelerate research • Engaging the public in dialogic archiving, where new information informs the catalog, and the collection informs cultural renewal (cf PARADISEC) • lack of time • evaluating expertise in reproducibility • No collective definition of what it is to be indigenous in administrative data • Inability to study mobile devices and mobile app use across a variety of devices • multi-org shared social sciences data deposits with multi-org governance • Access to integrated infrastructure, cost barriers, skills barriers etc • more opportunities for testing ideas fthrough field experiments • Tracking individuals as they age to ceate cohorts at community levels to capture economic & social impacs as they age and progress through life • time and resources • we have very good integrated data assets, but they are hard to use. We need to make key datasets like PLIDA more informative to researchers who want to use them. codebooks , disctionaires, classifications etc. • Indigenous data governance • Opportunities for using admin and related data to identify opportunities for running field experiments and then undertaking the field experiments • long term storage of data with appropriate governance • access (i.e., API) • Data governance policies and documents can be shared among many projects, less wheel reinventing • Having a prioritised list of data we need to integrate for research (+3) • An university approach • Give Aboriginal & Torres Strait Islander People control over their data (data about them) - which will allow them to decolonize the data & provide important context to data

QUESTIONS	RESPONSES
	• Lack of opportunities to learn from other disciplines how they are innovating in ways we could try. • methods of sharing data • Infrastructure partners & support (+1) • Standards based interoperable repositories (models and implementations) Language Data Commons of Australia • Integrated and connected data providing a more comprehensive priacture of pepole in their social context • decision ready modelling, model standards, model sharing, deployment and availability • community/NGO data to draw from? • It's difficult finding the right information as there are mnay entry points • UNSW Align • Australian Digital Observatory • digital literacy and AI workshops • Decadal Plan implementation • Australian Internet Observatory • gaining benefits from AI • focus on equity • Improving Indigenous Research Capabilities project • RI business and operating model • Australian Urban Observatory • The Australian Government Data Catalogue is being built to become a central catalogue for users use if they're not sure what data is available and whre • I am currently testing the ability to align research and infrastructure data management with Strategic Impact to create greater visibility and outcomes for our organisation. There are multiple projects across the organisation who are also testing different use cases. • IRISS outcomes • Indigenous data governance framework • Data Science, Advanced Computing, Skills Development, Software Solutions/ Platforms/Enterprise • Digital infrastructure hosting & support • DMRC Computational Lab, and QUT Generative AI Lab

QUESTIONS	RESPONSES
	• Australian Data Archive - social science data repository since 1981 • Metadata and data standards • The Australian Digital Observatory ARDC Funded • Existing national data assets - research (AES, AUSSA) and Govt (HILDA/LSAC, ABS, PLIDA) • existing pilot activity on entrenched disadvantage • The pilot underway in HASS&I can be broadened • 0-5 years data • Repositories and workspaces program • Indigenous data discovery and catalogs - • Digital infrastructure Cybersecurity • Various datas storage facilities well established • Gov versus Private misalignment - need to test the 'work' and how it aligns for this to be realistic. Ladscape scan to understand the dynamic ecosystem (industry/gov/private/universities) • Partial coverage of life course using admin data • pool of researchers, data scientists, reseach assistants in a university department focused on research & engagement (no teaching) • Geospatial, liveability data • Potential to link to Breakthrough Victoria Funding • Data catalogues • Working with manuscripts to make them into textual resources (Nyingarn) • Using data for policy • System architectural standards • Information in Society EIP - research network and funding • Metadata know-how • ABARES (Dept Agricuture) data tools • A secure data lab that is enabled to create shared environments that permit the sharing of code, curated data, programs for analysis, memos, etc. and that can house commonwealth data as well as other data (university infrastructure) • Community Data Lab • Australian Text Analytics Platform • stakeholder/public engagement

QUESTIONS	RESPONSES
	• TLCMap. When consulting researchers across disciplines, we found mapping is a common need. TLCMap needs just a little more work to become a sector standard, and collab closely with GLAM institutions. • ABS Life COUrse Data Assett • improving metadata • copyright and Indigenous IP • https://ladal.edu.au • The opportunity cost : Beagrie, Neil, and John Houghton. 2014. The value and impact of data sharing and curation a synthesis of three recent studies of UK research data centres. Report.http://repository.jisc.ac.uk/5568/. • Digital media platform data • We are focusing on bringing in NGO/Health data into national integrated data assets • creating metadata catalogs - • Full time engineers, data visualisation experts, security services, virtual environments, pool of research assistants, • AURIN spatial and digital twin assets • Connectivity and advocacy back to NCI • Metadata Aggregation • The ONDC has 3 main programs to help improve transparency and reusability of Australian Government data: * Data Inventory Pilot Program *Australian Government Data Catalogue *Dataplace • Documenting datasets we used in our research (client owned) • Data donation models • Interdisciplinary engagement • Convening research expertise • Disseminating key messages to social science research sector • Various metadata catalogues well established • Access costs for administrative data • access to data • costs of accessing data • access protocols • The ONDC has developed Dataplace to try make it easier to share data from some Commonwealth entities

QUESTIONS	RESPONSES
	• funder data policies (eg re: data access) • Lack of research software careers, policies, training. • Develop research software career pathways, policies, and training initiatives in alignment to / collaboration with national and international RSE initiatives. • Research software careers, policies, design & engineering models & operational processes, training. • create soc sci infrastructure specialist career pathways • https://society-rse.org/; eResearch Australasia; various Australian STEM research software teams / infrastructures / initiatives. • social sciences technical (infrastructure) skillsets (developers, navigators, connectors) • ResBaz multidisciplinary tech events to introduce social scientists to new infrastructure • abiding to ethical practices • ethics infrastructure • ethics specification for data use - defined by communities - but implemented by some combination of researchers and institutions(• Showing the value of what you are doing outweighs the risk (to the public) • engaging the public in the so what? • Australian public goes 'wow' about what social sciences contribute (virtuous cycle) • Making things with public AND scholarly appeal and uses – the better to secure support • Having social license • public engagement • Address societal challenges housing, climate adaptation, mental health • Data to allow reporting on national and international social targets • More targeted and timely policy insights • Data that supports actually meeting the Closing the Gap targets (+1) • opportunity to tackle the wicked problems • improve the life of Australians:fairer, healthier, more prosperous • Address structural imbalances in public knowledge, eg women's sport

QUESTIONS	RESPONSES
	• demonstrate impact on real world problems - has to be part of the program of work (+1) • digital driven intervention at scale: eg: social science driven interventions at scale in response to live data (+1) • High quality program evaluation that leads to more efficient govt expenditures toward interventions that work • We don't feedback benefits and impact of our research to the organisations that create or manage the infrastructure, this makes it harder for them to keep the infrastructure funded • Ability to help people understand their value and position in society, and show how they matter • understand how to shpae advancing technological to support our priority social outcomes • build socially-positive or socially value-adding tech (e.g. alt social media) • Reliably inform public and political debate on important and controversial issues. • digital literacy - know how's of using technology capabilities • New and expanded educational resources (schools and HE) • upskillling understanding and capability inuse of digital data resources • Universities paying researchers to run the training that social scientists need. Sometimes it's better to learn from your colleauges than a vendor representative. • social science appropriate data and digital training • undersgrad and postgrad training that is fit for purpose and prepares students for future research careers • Culturally appropriate explainer videos on building technical capacities on collecting, using & sharing data [via multiple platforms] • Training and development • Mapping all the offerings of training available free and otherwise with regards to researcher conduct and practices -to reduce confusion for all • DReSA training catalogue can expose current training offerings • Training/building capacity for new data users

QUESTIONS	RESPONSES
	• Leverage national communities of trainers eg Carpentreis instructors • Understand people in context by linking individual and contextual data • Understand complexity through linked data set analysis • lack of linkage and integration among datasets • unlinked data – how to connect? • Collecting data from public, private, and other ways to permit linking at different levels (person, family, geography, group) • data linkage and sharing data • expertise ianalysing linked data • data linkage/analysis in soc sci expertise • Risk of data breach • Operation and cost of sensitive data environments • What is the appetite for synthetic data that is hyper sensitive? eg. Is there merit in creating synthetic data and fuzzy digital twins? • MIDL, shared data environment for the curation of sensitive data, especially sensitive data at individual, household, group, community level • Build a Digital Twin of the Australian Society / economy / .... (+1) • disparate RIs • awareness of existing capabilities that can be leveraged for social sciences research • facilitating cooperation among people building the same stuff (+1) • knowledge about existing RIs • knowledge of how to engage with these RIs • Ongoing maintenance so the grant investment is not lost when system is launched at end of project, but there's no money or staff if it breaks. (+3) • efficient use of resources • operation and cost in org. silos. (+1) • Researchers actually using the infrastructure • Emphasise the cost of not investing in research infrastructure • getting more researchers to use our best tools (e.g. PLIDA) (+1) • Reuse of software and data, so you don't have to do it all over again, and can instead collectively build on systems and data. (+1)

QUESTIONS	RESPONSES
	• Digital Research Infrastructure build (ecosystem of resources - social media databanks, tooling, analytics, workshops), maintenance, and operation • "TLCMap. When consulting researchers across disciplines, we found mapping is a common need..." - "Why not use existing open source maps? Why use a fragile tool that may disappear when funding ends?"- “Many reasons. They are mostly for science and commerce not HASS. You can export all the work you put in to TLCMap, so it's not a waste of effort. But yes, we just need that bit of extra work to make it something that doesn't dissappear when funding ends, which is what we hope to get out of this current process. I could give plenty more reasons. but won't waffle. But eg you can create a standalone single map that dissappears after funding ends, but with TLCMap it is part of a broaders system that enables you to compare and contrast with all other maps in the system. To search within and across layers. And it's easy to just get a map up and online without a 6 month funded project. Etc. And you can export in a standard ready for archive in an archive for longevity. The point of this decadal plan is to find a way to make systems, not dissappear and stop the waste of everyone's time, money and effort. To work together on systems that are useful for everyone, filling research needs not provided by off the shelf stuff. This is the exact situation TLCMap is in. So if ARDC can't solve this problem for TLCMap and the many systems like it, that are real and ready right now, there's no hope for any, and ARDC itself will be just another system that dissappeared when the money ran out."

Activity 2: Opportunities identified by breakout groups

GROUP	RESPONSES
Group 1	• Cradle to grave data trajectories, with an ability to drill into community level data. Combining govt, NGO, research data • National longitudinal study of older for 65+ • Upskilling the social science research community. Pathways for career progression. Enabling discovery and access of existing research infrastructure • Training materials around Indigenous data practice. Capacity building, data management skills, training materials, Indigenous data governance, and cultural competency for non-Indigenous
Group 2	• workforce, skills, capabilities • project to cluster capabilities • service marketplace
Group 3	• Scaling and sustainment (governance) of context aware access to data • Lifecycle management of data from grant submission to archive • Policy modelling/evaluation functionality that can measure policy impact (Impact narrative demonstration) Link this to a grand challenge (e.g., Closing the Gap)
Group 4	• Training -understanding requirements, connectedness in terms of training, coherence, need for training in methodology - currently: tendency for training siloed • National guidelines and org guidelines not quite enough, ARDC guidelines could be useful esp in the indigenous space, Could look what's currently available to develop training resources. • mapping of the research cycle and data cycle and ethics cycle for each project in the form of plan
Group 5	• Indigenous data governance framework(s) • Beneficial across all areas of HASSI, and across organisations. A Key competency that is lacking in some areas." • A lack of domain specific knowledge that overlap with technical skills. Payment in the sector is also lower than other options. Could there be a pool of RSE's that are available?

GROUP	RESPONSES
	• Equity in public knowledge. Much of our data isn't accessible for everyday people. How do we make it accessible for the public to contribute and access?
Group 6	• Valuable contribution of social scientists can be maintained by having the tools needed to do the job. • Opportunity to create reliable information to inform public and political debate. • Infrastructure continuity. What does infrastructure look like when it is not a series of projects but a continuing ecosystem? How can continuity of staff, systems and data be built into anything created in next 4 years. • Employing workforce under conditions that lead to staff stability and ability to keep working on the project beyond the life of project.
Group 7	• General lack of resources : time/costs/people • Awareness of reuse potential across multiple dimensions (Policy, practice, activity) as well as data and compute • Skill mobility across institutions and/or project activity • Bonus challenge. An infrastructure ecology sustained by benefit
Group 8	• Knowing what data assets already exist (data discoverability); a lot of data is still siloed, insufficient metadata. Connecting researchers at universities with government departments and data. • Limited funding and resources (cost is a barrier). Access cost for administrative data. • Training and capacity building • Enforce public availability of publicly-funded research and research outputs (including data)
Group 9	• Data asset quality (national standards for data collection; metadata about the collections; better linked assets) • Data asset findability • Resources (cash, human resources) to work on data • Also skilled workforce -> incl training/certification/recognition • Arrangements for access to data assets (costs, turnaround time, access arrangements, governance) • Building social licence; assets of value to the community as well as research (potentially as a dashboard?) -> see Uni Melb - 'breaking down barriers' project as an example -> ie. data asset of high value

GROUP	RESPONSES
	to researchers that they can utilise, but that also has a dashboard/UI that the public can access to find information about their suburb/demographic (without compromising the research usage) so it has value to them and they support the research use by proxy • Outcome: social science is considered a 'real' science (like medical science etc)

Activity 3: Reflection

QUESTION	RESPONSES
What was the most exciting opportunity you heard today?	• Scalable and sustainable governance over access to data • Things that interoperate with other research infrastructure. • Data discovery and governance • we actually talked about problems. These were depressing • Having established frameworks and structures • A pool of HASSI aware RSE's ! • The commitment of social scientists to commit to making a positive difference & need for infrastructure to maintain this. • coordinated access to data • The idea of an Australian version of the US Health and Retirement Study • Lifecourse data trajectories at community level • The cradle to grave data trajectories. • Workforce - livable wages, career pathways, skilling, planning past project lifecycle • Thinking about staffing and supporting social sciences digital projects in a way that can make outputs greater than the sum of their parts • Discovery and access to existing 800 RIs; upskilling and career path for RI staff • Need to demonstrate impact on a real world issue in 4 years • Coordinated approach to enabling access to sensitive data (governance, tech and infrrastructure) • Not 1000s of projects, but infrastructure we can all build on and benefit from.

QUESTION	RESPONSES
	• that perhaps teaching people to share might be a good start? • connecting the 800+ assets into something bigger • Cradle to grave data cohort trajectories • further enhance some of our key national data assets for researcher use • Project to bring HILDA into PLIDA • Acceptance of technical/computational skills as being necessary for the advancement of social sciences • Training materials (non-traditional formats including short culturally appropriate videos by local Indigenous "data liaisons") around Indigenous data & technical capabilities development. • UG/PG research training skills improved/extended to anticipate the full data life cycle from questions, methods, data collection/reuse, curation for reuse and or archiving long term • Shift from a discussion of projects to persistent, sustainable infrastructure • Indigenous research capability uplift • Demonstrate functionality in policy evaluation • The idea of creating a dataset with broad applications so that communities can access dashboards and build social licence • Building collaboration across the social sciences and raising profile of value and purpose of social science expertise • DataLab enhancements (UI, analytic and data visualisation tools etc) • Focus on workforce needed - how to network so they are not isolated • national whole of population data asset with relationship links across different units, such as persons, households and families/generations • Recognition of the role of social infrastructure surrounding in-demand expertise.

APPENDIX 2: SURVEY RESPONSES

QUESTION	RESPONSES
What data or digital research infrastructure do you currently use in your work relating to the challenge of building a more equitable and resilient society?	• ABS data lab. administrative datasets. High performance computers (sometimes) • Not sure. • Social network analysis (SNA) is my main area, which focuses on the patterns and implications of social connections. As a relational methodology and theoretical framework, it highlights the impact and importance that our social connections to others may have for us individually (e.g., feeling mentally well due to having friends) and collectively (e.g., sharing knowledge effectively to develop new innovations, like a COVID vaccine). • A broad suite of spatial [temporal] data sources • University on-prem, Azure. • None yet, but in the future I will • Nothing in particular but use the media and inquiries to highlight the consequences of poor statistical work eg Robodebt • integrated administrative data (PLIDA, state govt data), longitudinal surveys (HILDA, LSAC, LSAY, LSIC); surveys of higher education students and graduates (QILT) • Linked data that incorporates routinely reported national data, health sector-derived electronic record data (a national first), environmental and geospatial data, supported by both traditional analytical systems and new capability in artificial intelligence techniques to mine the vast capability of unstructured data resources. • several longitudinal cohort studies and administrative data • I use a variety of tools and infrastructure: local computer (with Stata, Python, R, QGIS, etc), virtual machines (running on Monash SeRP, on AWS, on GCP), the Data Basis platform (at https://basedosdados.org/, with harmonized public datasets). • Trove/National Library of Australia; National Archives of Australia; Parliament of Australia; ABS. • R software and packages, mostly. • PLIDA, HILDA, service data from a not-for-profit organisation

QUESTION	RESPONSES
	• Geospatial Information Systems (GIS) to analyze spatial data related to regional development, environmental management, and community planning. GIS allows researchers to map and visualise various social, economic, and environmental factors, facilitating evidence-based decision-making processes. Data Analytics Platforms: to analyse large datasets from diverse sources, including social media, government databases, and sensor networks. These platforms enable researchers to extract valuable insights and identify trends, patterns, and correlations relevant to promoting equity and resilience. Digital Repositories and Archives: to store and manage research data, documents, and multimedia content. These repositories provide open access to valuable resources, fostering knowledge sharing and collaboration among researchers, policymakers, practitioners, and communities. Community Engagement Platforms: to facilitate community engagement and participation in research and decision-making processes. These platforms enable communities to contribute data, share their perspectives, and collaborate with researchers and other stakeholders in addressing local challenges and promoting resilience. Online Learning and Capacity Building Resources: to build digital literacy and skills among various stakeholders, including researchers, policymakers, practitioners, and community members. These resources empower individuals and organisations to leverage digital technologies effectively in promoting equity and resilience. Interdisciplinary Collaboration Tools: to facilitate interdisciplinary collaboration among researchers, policymakers, practitioners, and community members. These tools enable real-time communication, document sharing, and project management, fostering synergy and innovation in addressing complex societal challenges. • A wide range of data from Trove news articles, to ABS statistics, to historical data and maps that we digitise. • I use collection catalogues at a university that are not interoperable. They are not easily accessible to people who aren't connected to the Uni library for eg. Some data is not findable.

QUESTION	RESPONSES
	• [University] Data Lab, ABS DataLab • PLIDA, ABS DataLab, and PHRN state data linkage nodes. And the Internet! • We use tertiary institution libraries; internet data sources for instance Population Reference Bureau - USA; latest library primary and secondary sources; presentations from conferences thematically related. Probably the whole issue is extensively fragmented, with only libraries and internet search engines remaining as the most reliable. • Qualitative researcher so focus on situated experience of systems; often the missing perspective. • PLIDA, DEX, BLADE, HILDA, LSAC, LSIC, BNLA + international longitudinal data such as Understanding Society (UK), GSOEP (German Socio Economic Panel). • I employ computational infrastructures that facilitate the development of my toolkit as well as the collection, storage, and analysis of large social media datasets. I utilise Nectar computing resources, including custom instances and the Virtual Desktop Service, to support these tasks. Moreover, I employ advanced machine learning techniques, including large language models like GPT, to enhance the depth and efficiency of data analysis. • Various open government data repositories (U.S., Australia, UK, EU); Wikimedia & Wikidata; persistent URLs (PURLs) for digital library infrastructure.
What do you wish that you were able to do in your work towards this challenge? What would you need to be able to do it? Consider opportunities to enhance or build on existing research infrastructure, to	• Access admin data that tracked people from infant to adult (age 30+). • Upscaling of project knowledge and National interconnectivity would be very useful. • I have had to build my own social network software platform in order to do things like collect network data, visualise it, analyse and report on it. This has taken me significant time and effort. Had this infrastructure been in place, I would have been able to more easily and effectively conducted my research and given feedback in a timely (and thus useful) way to partner organisations. • Seamless access to linked, harmonised, longitudinal spatial data describing a range of social phenomenon

QUESTION	RESPONSES
connect and support interoperability between existing research infrastructures, and to develop new capabilities that add to and support existing research capabilities in the broader research infrastructure ecosystem.	• Developing a HASS digital research hub capable of supporting local but also national grand challenges questions, enabling computational analysis and data management at all scales, and implementing best in class research software engineering methods, management, and career paths for that purpose. Alignment to and integration with national RI is crucial to that goal. • I would need to be able to access data for people who identify as trans - can be sensitive data • Improved documentation and metadata for administrative data • Integration of the main Australian longitudinal surveys (HILDA, LSAC, LSAY,) with PLIDA • Integration between state and commonwealth admin data • Integration between higher education admin data and QILT surveys of university students and graduates • Faster approval, processing times and lowers costs associated with DataLab access • We would like to build on our proof-of-concept studies and build novel and large scale capability to identify social vulnerability in the population, and its impact on health and well-being. Towards this we envisage that we will need: 1. Create new datasets indicating social and economic vulnerability based on data sources that have not been used before - eg unstructured health sector data. 2. Establish new linkages between our rich health and community data with national datasets of social vulnerability (eg disability, income, built environment and housing) 3. Expand existing data linkages with federal and state datasets 4 Develop and embed new tools for capture of features of social and economic vulnerability using artificial intelligence techniques (natural language processing) - we have proof of concept studies completed. 5. Build on our current plans to develop a policy simulation platform using large scale real data - targeting social vulnerability 6. Ensure a framework of responsible use of data access, governance and AI use, in tandem with community consultation and co-design at scale.

QUESTION	RESPONSES
	• A longitudinal ageing study for Australia - Have access to identified data through a secure virtual environment that is accessible online (e.g. via VPN) - Be able to do data in and data out of this protected environment, to merge external data. - Have the identified data be continuously cleaned and improved based on researchers' past work and feedback, so as to avoid every new user having to reinvent the wheel." • The NAA and NLA - including Trove - have experienced chronic under-funding. If they were better resourced for access, including digital, they would greatly enhance historical research. For instance, there are still few digitally available newspapers for the period 1955-95. Public access to born-electronic records at the NAA remains at a rudimentary stage. • More time and resources to improve the quality of documentation to make my software tools more accessible to users. • Linking additional educational data, such as University students not currently included in the spine because they do not receive HELP loans; linking additional educational information, including NAPLAN data and ATAR scores, including scores for those who do not 'use' them for university by being admitted via ATAR pathways. These are crucial to understanding causes and effects of educational inequalities and inequalities by socioeconomic background. • Improving PLIDA metadata and providing an understanding of how data are collected at the source. Often, due to changes in eligibility or policy over time, it can be unclear who is in a given data set or what we can infer about them if they receive a certain payment or report a certain level of income, etc. Working within DataLab as the only means of accessing metadata including data dictionaries and codebooks is also clunky, as the DataLab spreadsheets are often slow, laggy, etc. • Have a full time job rather than juggling 3 to five short time part time contracts and casual appointments. As well as alleviating the mental health problems the constant fear of losing my famil's home and potentially my family due to the relationship and mental stress that causes, and the ability to concentrate on doing a good job

QUESTION	RESPONSES
	instead of looking for more work when I already have too much, I could plan for longevity and sustainability of software better. Nothing else matters as much as addressing counter productive staffing. Staff continuity would mean millions of investment in projects are not wasted after short term funding to set up software systems runs out, and that instead of paying for months of work to upskill and onboard IT businesses or new temporary staff, while systems are down, issues could be fixed in minutes. This would also alleviate resentment that technical staff may feel towards academic staff being promoted off their work, sometimes without having contributed much, while they are rewarded with unemployment after every success. ● I would like to show Indigenous people how to use a connected collection ecosystem with ease of access. ● Get easier access to build and join together public and private sector data to study a range of questions and to enable better analysis, increased policy engagement, testing of ideas, and simply making a difference. ● I'd like for it to be possible to evaluate any and every Commonwealth or State social or health policy via an agreed minimum dataset for clients and programs collected by all service provision agencies (Govt or NGO) linked to other relevant data resources. To do this we require support to enable a minimum dataset, and elimination of barriers to State-Commonwealth linkage. ● Upskilling of social science researchers and increasing of software engineers in social sciences for guaranteeing progression in teaching and research and in turn innovation remains crucial. Researchers need extensive training in user-friendly software for social sciences progress and adaptation to challenges prompted by the digital technology wave otherwise the risk of lagging behind would be too immense to recover from. Higher utility of InterActive Software and SPSS by trained social scientists/researchers/ lecturers/ admins/ assistants/ (schools, tertiary, research; primary - secondary schools; computer use learning courses on cite) among many would be much consistent with the value generated by the envisaged AIESS

QUESTION	RESPONSES
	and also in other nations abroad. AIESS aims and mentors us all on systematic mobilization and maximum utilization of social sciences value on intra-national terms while creating a visible competitive edge that allows the same AIESS to derive maximum value from other international social sciences circles. This is the scenario that Zimbabwe, Southern and Sub Saharan Africa should prepare to now that we have this greatly enriching, empowering engagement with the Australian Academy. ● More accurate, at scale, transcription of audio would be helpful. Still not accurate enough for unsupervised application despite massive advances ● Access more lined administrative data in a timely and economically feasible way. I'd love to see more longitudinal data, such as HILDA and LSAC, linked to PLIDA. ● The infrastructures mentioned have significantly enhanced my research capabilities; however, data privacy challenges persist. To provide some context, there are two primary types of infrastructure commonly used by researchers in computational work. The first, akin to Google's Colab, allows immediate code execution without additional setup. This setup is convenient for researchers to build, test, and run code continuously without interruptions. The second type involves capabilities of large language models (LLMs), like GPT, which facilitate large-scale data analysis without the need to independently develop complex machine learning models. While tools like Colab and GPT are highly useful and convenient, they also pose privacy risks, especially when handling sensitive data, raising ethical concerns when such data is uploaded to these platforms. Although Nectar offers some excellent features, it remains challenging for researchers with less technical expertise. Therefore, I have two main suggestions for improvement: firstly, the ARDC could enhance tool accessibility in Nectar, similar to Colab, simplifying the process of developing or accessing multiple virtual desktops. Secondly, developing LLMs like GPT specifically for researchers from Australian universities would help reduce data privacy concerns associated with using these tools.

QUESTION	RESPONSES
	• I would like to co-create explainer videos and accessible training resources to build technical capacities in a culturally appropriate manner for Indigenous communities, academic researchers (HASS & natural sciences).
If you have any thoughts about these potential opportunities, or other ideas that you would like to raise, please include them below.	• In education space, first question is what work has been done on life course data trajectories. • Agree that the above would be very useful. Particularly climate justice perspective into life course trajectory and policy modelling. • I would dearly love to see relational data opportunities, such as social networks or other relational datasets, because the social is, by definition, relational. As a society, we need to understand how we connect (or do not). Asking people their individual opinions about 'how connected you feel' (or similar such questions) is a bad proxy for understanding the specific structures of social relations and their broader implications. Doing social research which uses only non-relational data and statistical methods that assume the independence of observations (i.e., one person's response is totally independent of another's) makes no sense to social, relational data where the connections represent interdependence. We need infrastructure support for this sort of data and analysis, and this can augment all of the above. • A single national longitudinal data asset of household travel surveys linked to other relevant unit record data • Lifecourse data trajectories would provide great opportunities • On the first option, you should collaborate with the ABS on what can be achieved through the longitudinal Census data base. To have maximum impact, policy modelling should be done in collaboration with key Government agencies such as Treasury, Productivity Commission and the Department of Finance. • Integrating major longitudinal surveys, including HILDA, with PLIDA is important, however the feasibility would depend on commitment from the key stakeholders (data custodians) • A lack of a longitudinal survey of ageing/older adults is currently a gap in Australia. I would start earlier than 65, at 50, to align with major international surveys and to enable studying transitions into retirement

QUESTION	RESPONSES
	• Integrating data of different types - such as area-based/spatially structured data with individual data - and developing tools to facilitate that, like the GeoSocial data integration service pilot, would be another fruitful avenue to pursue • Working with the ABS around improving the quality and useability of the PLIDA data and the associated metadata/documentation is another key area to pursue • Working with the ABS to better integrate data on individuals in PLIDA with data on families (over time) and communities where people live • My organisation is investing in the Longitudinal older adults data option - we have scoped this and conducted a pilot study and prepared submissions on it. • We have developed technology to harmonize hundreds of large public datasets and make them available through a search engine and data lake on Google BigQuery. It is quite powerful, and is about to receive 3 million queries this year alone. Integrating some of this technology to this opportunity could be fruitful. • I wish there was a portal where researchers could access private data but through a layer of 'differential privacy' applied. This would allow universal access to the data (with no project submission, etc). • The Academy needs a broad sense of what 'research' and 'data' might comprise. It isn't just quantitative. It isn't just applied, or policy-related. One of our grand challenges is to preserve and develop a functioning democracy based on a sense of inclusion and this will draw on a much broader sense than implied above of what matters in terms of both data access and funded social science research. • The above options are outside of my area of expertise. • They all sound important, particularly working toward lifecourse data trajectories to understand processes that take time and may occur across generations, e.g., intergenerational income and wealth inequalities and social mobility. Linking to additional, non-Commonwealth-held administrative data sets or other data sources would also improve the ability to understand these processes.

QUESTION	RESPONSES
	• In particular to the lifecourse data trajectories and longitudinal older adults data projects. • I am a proponent of the lifecourse data trajectories as I think it has the most promise for promoting ambitious (but achievable) work that will lead to great impact. The connecting HILDA and PLIDA is already underway and being explored. I agree that ageing is an issue -- but we're already working on a number of data assets (in Australia) that will enable more work on this issue. • All of these are important and useful. I can see practical issues with retrospective consent for HILDA linkage, and sample loss from non-consent to link at future waves. As for Lifecourse Data, the post birth- to 5 years space is in critical need of funding for national digitization and linkage of "Child Development" books via community child health nurse programs. • The strategy is a good start, but would encourage ASSA to consider the lessons of other jurisdictions and disciplines. The best comment of the workshop was that others have gone before us and we should learn their lessons before committing. • On the notion of grand challenges. I am all for a greater alignment with defined strategic priorities that both government and business are aiming for (as this solidifies funding, and opportunities for collaboration). But I'd be cautious as well. ANU and CSIRO have both walked back their "grand challenge" frameworks, and we risk an "orphan idea" if we get too caught up in this. • On the specific challenge suggested - resilience is a neo-liberalised instrument that implies responsibility and moral relations. Strongly dislike this. Wellbeing would be better (stealing this from Mark Crosweller recent speech where he said what we've all been thinking in critical sociology on this but had the guts to say it in government). • I'd be keen to support all of these ideas - they are all excellent. • Of the options suggested during the first workshop, I see the most alignment with the first option ('Lifecourse data trajectories') which as I understand it would support the government's National Agreement on Closing the Gap & Priority Reform #4. My second strong favourite is to support Policy Modeling, specifically public

QUESTION	RESPONSES
	health policy modelling because we do not have a resilient data ecosystem in Australia based on research I and others have performed post-COVID-19 response in Australia.

APPENDIX 3: WORKSHOP 2 RESPONSES

Activity 1: Benefits

SUGGESTED DIRECTION	RESPONSES
Improving data discoverability (for instance, by developing a "one-stop-shop" for initial information on different types of data available for social sciences research)	• my research in mobility and migration would benefit from: a one-stop shop of available [linked] data resources • well identified repositories - the one stop shop idea is very attractive if well configured • having a one-stop catalog of available data that is searchable • Centralised data to link multiple repositories, e.g. museums, archives, libraries, scientific reports (objects, documents, records, data) • With unified authentication and authorisation data could be federated - single place to search hooked into multiple catalogs/APIs • the need for national data integrating capabilities/authorities • Ensure existing data assets are made use of • better Fairness of data • visibility, awareness, and knowledge of data sharing policies and procedures • I think this would benefit PhD students/ECRs across the disciplines of social sciences - particularly those ECRs/PhDs/postdocs who come from international backgrounds and might not know much about Australian data - ability to search for relevant content would help them get started on their research • could seed new projects as people serendipitously discover different type so data. • More mixed methods projects as qualitative researchers felt more comfortable accessing data sets • The use of artificial intelligence techniques to capture key concepts and features in unstructured data • Removing technical barriers to entry would support a much broader range of disciplines engaging with these assets • Drastically decrease the discovery time, especially for grad students. • Saving researcher time searching for data in multiple different locations

SUGGESTED DIRECTION	RESPONSES
Enhancing the accessibility/ usability of administrative data (by improving the documentation and metadata for key administrative data assets such as the ABS's PLIDA)	• Multiple database viewer (across multiple databases/ institutions) • Consider use of the Data Availability and Transparency Act 2022 to support sharing of Australian Government Data. Research purposes included under section 15. • Dataplace can be used to request Australian Government data • It is not just about sharing of documentation, etc. but also enabling shared access • Can be used to make informed decisions regarding government policies and programs, R&D • Ensure all data made accessible adheres to agree standards • If this is not possible due to a lack of available information, make a set of tools available that allows users to easily curate data themselves • some PLIDA modules would benefit from enhanced documentation and metadata, e.g. DOMINO/DSS data, or health data - these are often not intuiitive dtasets particularly for somebody who is not very familiar with a particular module (e.g. education researcher who also wants to include health-related variabels in the model will likely struggle whit understanding the health data in PLIDA) (+1) • the value of lowering the barriers to access multiple [protected] data sources • Access should be considered with regards to the context and privacy rules • sociology geography; urban planning; environmental science; criminology; demography; political science; economics • Code libraries and analysis handbooks (with examples) will make all projects more efficient, huge duplication of effort currently in analysis of routinely collected health data (esp pre-processing) (+1) • it is rare that a given admin data set can be assessed, curated and ready for use for all types of questions. Therefore, being able to access information on challenges faced by other analysts and how they resolved the data challenges would progress my ability to work with a data set confidently and with greater speed (from access to analaysis)

SUGGESTED DIRECTION	RESPONSES
	• Better documentation will help researchers decide if a dataset is suitable for their project before applying for access.
Further integration of (individual-level) data on people (for instance, by integrating major longitudinal studies with PLIDA)	• We could better make the case for poverty alleviation - cross data linkages will clearly show this • This is already being done and should be a part of the fabric for continuous improvement • the value of pre-linked unit record data that has been linked via repeatable/robust and transparent analytic processes is essential to ensure the credibility and quality of research data • Individual permissions to data • Enhances ability to do deeper dive into analysis, both in terms of the connecting across areas (e.g. role of schooling on enabling opportunities for reducing poverty), and to permit deeper dives within groups - could be an age cohort, a geographic location, or some other grouping • Identification of particular population subgroups is v challenging using administrative data but is made easier when data are integrated (e.g, identifying those experiencing insecure housing) • Streamlined access is key to ensure researchers can access the data when they need it • Requires strong relationships with data owners and trusted organisations to govern use • Linking PLIDA with other forms of admin data (e.g., state education data, private-sector housing data, non-profit service provision data) would open up a range of new research questions across those domains and allow for better understanding of causation. Eg, what factors are correlated with low-performing primary students (low NAPLAN scores) who are able to succeed later in life (perhaps using adult income from PLIDA as an outcome)? • Being better able to control for family background and SES on a range of life outcomes, including health, labor market outcomes, family formation, etc. • Bringing together state and commonwealth educational data would offer a tremendous potential - e.g. ability to study trajectories through school and into post-secondary education (VET/HE) and the

SUGGESTED DIRECTION	RESPONSES
	labour market. This is currently not possible but good models exist - e..g. linkages in the health space (linked state-based hospital data and commonwealth-level PBS/MBS data) ● Intergenerational studies if we link families. ● Does the neighborhood you are born in determine your lifetime earnings? ● while linkage and integration of these datasets can be an aspiration, it is essential to consider the privacy implications and ethical considerations up-front to pre-empt negative un-intended consequences ● Would support greater insight into inequalities across domains (e.g, health as a driver of income inequality, income as a driver of health inequality) which are currently quite challenging to disentangle empirically ● enable interdisciplinary research ● sociology geography; urban planning; environmental science; criminology; demography; political science; economics ● cross-sector and cross disciplinary data integration is of highest value I would think
Cross-unit data integration (bringing together data across different units, such as individual, family/household, community, etc, including spatial data)	● Consider cross linking social and health/environmental administrative datasets for individuals in geographies - this would be a powerful resource for all types of researchers including social sciences ● The value of spatially integrated social sciences talks directly to cross unit data integration [alongside a cross-cutting theme cutting across the other domains ● enable intergenerational or other, e.g community level analyses of person-level processes ad outcomes ● Cross cultural ● this component lies at the core of my research program - the need for robust/transparent and repeatable analytic procedures to join data across units is essential to address complex social science questions ● the capacity to map, measure and monitor phenomena over time is essential and where cross unit linkage is critical

SUGGESTED DIRECTION	RESPONSES
	• spatially integrated social science, reseearch questions at the intersection of social science and other disciplines • GeoSocial pilot could be extended to include broader range of surveys and broader range of spatially-structured data • getting all these data linked and integrated will require figuring out the data custodianship or stewardship issues as well as the technical or physical repository that is sustainable for a foreseeable future • con sider the advantages of integrating strong social sciences data with health sector data - something that would value add to both social sciences and health sector researchers to explore social determinants of health or vice versa in a really unique fashion. • new cross-domain projects - eg,, social and natural or built environment work for research & planning • sociology studies, cross unit analysis, policy impact assessment • sociology geography; urban planning; environmental science; criminology; demography; political science; economics • cross cultural adaptations of such data, migration, health, environment • Could include data from three levels of pre-tertiary education. • ABS is developing family and household structures for PLIDA but everyone's definition will be different. I guess that as long as we can make all the data assets are interoperable at one type of unit, particularly the person, then we should be ok
Connecting data over the lifecourse (for instance by connecting existing data sources about different groups of individuals to give a picture of life course trajectories)	• life is complex connecting aspects within and over time is important • Data repository maintained in perpetuity (e.g. museum database) • Would support more robust study of factors with very long lasting impacts, short studies may show "no difference" or "no impact" but significant differences (e.g., in health) emerge over subsequent years • Some data gaps need investment. For example HILDA does not include remote areas. Lack of good census data for housing and community planning: https://openresearch-repository.anu.edu.au/handle/1885/147833 • The kind of catalogue of resources that is mentioned under Theme 1 (discoverability) could be used to map out what is available in

SUGGESTED DIRECTION	RESPONSES
	which area/on which topic (e.g. health or educational poutcomes), and over which life stage. This could also be used as a 'gap-analysi' tool to inform new data collections where existing data might not be available (e.g. older age surveys) • Looking at the timelines for specific maltreatment impacts, e.g. not finishing school, then homelessness, then own children removed etc. • poverty is a complex issue .. no one solution -- accessing data on those on poverty exists, if i could observe those near poverty and understand dynamics related to family, communities, as well as own activities (schooling, use of services, health, etc.) then I would better be able to understand the importance of different types of potential interventions for addressing poverty • How does environmental degradation impact in-utero fetus health and later life outcomes? • Intergenerational trauma - tracking the impact of childhood trauma and the impact of interventions across the lifespan and next generation • Determine whether interventions are effective • would support cross-domain questions - e.g social, economic & health • disciplines that could make use: health policy, built environment: architecture, urban planners,urban • The true value of lifecourse data is in the combination of different types of data from cross-sector disciplines • sociology geography; urban planning; environmental science; criminology; demography; political science; economics • Criminology and Adult Ed. would benefit from this • A first stab at this could be connecting existing cross-sector data resources - health, income, environment, housing - these have relevance to life span outcomes for people, and could be done at cross-sectional decadal level • life courses of people, places and other 'things' - e.g. land uses/housing/vehicles

SUGGESTED DIRECTION	RESPONSES
Considering Indigenous data governance in the context of social science data (working out how to incorporate Indigenous data governance principles into integrated public sector administrative data)	• Centralized data repository with First Nations permissions attached which links / viewer to multiple databases. Allows 'Ask First' protocols. • Refer to https://www.pc.gov.au/inquiries/completed/closing-the-gap-review/report/closing-the-gap-review-supporting-paper.pdf • the need for far better understanding of indigenous data governance is critically needed - and now! • community involvement in research for community benefit • enhanced social licence for Indigenous research • contribute to a modernised social science that is participatory and socially engaged • this is important but is also important for all data sets, expecially any sensitive data • would support the implementation of CARE principles across the HASS-I RDC • I think integrating principles around Indigenous data governance for publicly available data would help to support Indigenous-focused research using these assets (e.g. PLIDA is in principle good for capturing Indigenous people's outcoems because of its size but currently no guidelines exsit for how wuch studies focused on Indigenous people's outcomes could be carried out using these datasets) • would support the development of Indigenous research capability • activities here need to be integrated and aligned with all the other activities about indigenous data governance that are already happening in other funded projects such as the Indigeneous Data Network, Language Data Commons of Australia, etc. (+2) • In the future there will be a National Data Catalogue which makes Commonwealth, States and Territories data to be discoverable • Work is being undertaken by to develop an Australian Government Data Catalogue to find open, conditional and restricted data that can be shared • Database of First Nations Ask First protocols for use by indigenous researchers working on other nations' country/

SUGGESTED DIRECTION	RESPONSES
	• Who do I ask for permission to access this data (First Nations) • Indigenous owned and controlled data sovereignty- i.e community controlled and accessible to community (+1) • Better understanding of place of Indigenous people in Australian society • sociology geography; urban planning; environmental science; criminology; demography; political science; economics

Activity 2: Requirements and enhancements

SUGGESTED DIRECTION	RESPONSES
Improving data discoverability (for instance, by developing a “one-stop-shop” for initial information on different types of data available for social sciences research)	• Consider contributing to the Australian Government Data Catalogue (+1) • AI techniques for discovering concepts and features not traditionally recorded using structured data • Data with breakdown by city would help in inter-urban studies of housing • Housing, Finance, Environment, Disability, etc - variables that people from health sector cannot usually access - to make them more discoverable, a "one" stop shop to find the breadth of cross-sector data for example to study determinants of healthy ageing • Comprehensive coverage in terms of topics/domains, should coverthe key existing survey and administrative data • Tools and training to be able to get the most out of the systems • The need for dedicated national data infrastructure that acts as the conduit to access data and the requisite embedded tools to procure the required data. • Metadata fields for datasets: how to apply for access, cost, past applications, etc. • Somme assessment of the coverage and quality of information (e.g. is this dataset adequate for studying mental health outocmes in children etc), and some assessment about the accessibility (e.g. restricted access etc)

SUGGESTED DIRECTION	RESPONSES
	• For concepts that may be measured by multiple sources in different ways, e.g., disability, being able to quickly understand differing definitions and all potential sources of those variables • A mapping of who is making what discoverable - there are many data catalogues and even many duplicates of the same data. Cordination of efforts is key to ensure the one-stop-shop is coherent and doesn't become a maze • Project to audit what data sources are being used by social scientists in Australia and for what • some indication of potential applications • Examples of previous research to demonstrate how the data were used before. • data asset quality assessment or indication of known issues or pitfalls • More ai language based models that could assist with searches across varying terminology • standard classifications • A powerful search engine (ElasticSearch, Solr) based on a meaningful metadata schema. (+3) • Ability to filter by domain but also by more specific outcome types, by the sateg of the life course, by the type of data (e.g. admin, suvrey) • having the relevant information but in a succinct manner and in a way that I can raise questions / query the information. • Multi-faceted search will be important e.g. show me only the data I can download right now, or more ambitious show me only the data that's compatible with my analysis workflows (+1) • Metadata fields for columns: coverage (spatial, temporal), type, description, measurement unit, foreign key. • Metadata fields for datasets: coverage (spatial, temporal), unit of observation, update frequency, automatic or manual, tags, themes, organization, etc. • standard metadata across data assets • Consider making metadata of more sensitive data to be discoverable and separating accessibility as a separate matter

SUGGESTED DIRECTION	RESPONSES
Enhancing the accessibility/ usability of administrative data (by improving the documentation and metadata for key administrative data assets such as the ABS's PLIDA)	• Again, the information about the data in a manner that is understandable across disciplines; health to social sciences, social sciences to health - for example • Victorian Social Investment Model (VicSIM) is very useful. A national version would be exceptional. • More robust governance agreements across the board • ? perhaps facilitating discovery of experts who have experience working with these data for research purposes. These technical folks are usually hidden within teams/institutions and finding them is difficult (but rewarding!) • governance standards, • Expert support & bringing people together • respect for data custodians (and subjects of the data), relevant information • Accessibility will require strong and trusted relationships with data owners and/or custodians. Again do we need a coordinated effort to collectively achieve this • standardised tools housed within a single environment supported by the requisite training materials and protocols to ensure safe/ethical use of the data • Training in working with analytic methods in the DataLab environment)as opposed to training around the access) • campain and researcher training to support re-use of standard data assets - not bespoke data linkage or new collection • Quantitative analysis is not always supported in the social sciences. Start with equipping students early with the skills. Social workers can significantly benefit from large data sets as it positions the individual within the system. • Capacity to store and process big data (on the order of hundreds of GB) • additional access opportunities outside a few secure analysis environments • Enhanced capacity to accredit appropriate, external analysis environments (e.g Monash's SeRP) would allow us to more efficiently combine administrative data with research data. Needing

SUGGESTED DIRECTION	RESPONSES
	to push everything constantly around between environments is a pain and causes huge delays! (+ governance headaches!) • considering data rights and data access management through a central point • indexing research outputs which have drawn upon these data would mean we don't have to each re-invent the (methodological) wheel in our own studies and grants. • central derivation of commonly required variables and hold these within or alongside the source data. In health, we do not need 100 teams trying to compute the same diagnostic status or outcome measure. (+1) • Mechanisms to share the code between researchers in the DataLab, plus some platform to enable shatring relevant publuications, icnlduing technical papers • Scripts to clean standard/commonly accessed data sets • indication of administrative "business rules" and how these influence things like variable definition - • Documenting and understanding definition shifts (in health) is critical. Changes to the clinical basis of code definitions, or funding eligibility can drive massive changes in the data which are not connected to changes in the experience or wellbeing of individuals • Improved description of the administrative process behind the data and how the data are collected. • Again coordination is critical - collation and curation of data will need to be a team effort and will therefore require agreed standards across all involved • Have a forum where users can freely discuss issues about the data. (+1) • More information about data limitations, especially the coverage. • standard metadata, classifications, definitions etc • Establishing consistent definitions of data collection categories (e.g. even the definition of 'child' varies from 16yrs to 18yrs) • understanding what data is available alongside what is and can and cannot be linked. Metadata to chart details and possibilities for linkage

SUGGESTED DIRECTION	RESPONSES
	• Bringing some of the PLIDA metadata and documentation outside the DataLab - so that potential users can browse through the content withou having to be trained and approved for Data:ab access • Knowing about a data set one one thing --- understanding how it can link to other data, and other critical information like period covered, population covered, etc. is super important • provide as much open information as possible about sensitive datasets to support discovery and re-use • Create placeholders for more challenging metadata • The ability to view metadata widely and in a user-friendly way via a public website, with more detailed/sensitive metadata available separately for approved users
Further integration of (individual-level) data on people (for instance, by integrating major longitudinal studies with PLIDA)	• Have a central 'data office team' + representatives in each data custodian organization that can interact and maintain the platform. • NIHSI AA data developed by AIHW in the health space could serve as a model for integration of state and commonwealth education data • Consider integration of existing cohort data that are designed for social and non-social sciences purposes to maximise benefit • Datasets cross referencing LGAs with categories of regional and rural areas would assist multidiscilpinary projects. • Social assets that are critical for health - people assets, money, education, employment, housing, accessibility to healthy options in life • Full educational histories • Need to cater to both sides, data owners want security, researchers want flexibility - we need a balancer in between • code libraries for things like data management, variable construuction, measure development • Training in using longitudinal data, partuicularly when intergrated with administrative data • tools such as the demonstrator developed as part of GeoSocial are essential to ensure consistent, repeatable and robust linkage - the pathway to high quality research data

SUGGESTED DIRECTION	RESPONSES
	• Critical is having a range of ways to access the data (not a single location) • Being able to utilize the data in shared environments to enable the accessibility and usability of admin data would be critical • Taking commonly required workflows and provide these as a service • making data re-usable using meta data standards and definitions
Cross-unit data integration (bringing together data across different units, such as individual, family/household, community, etc, including spatial data)	• support for exploratory discussions across disciplines to identify new cross-domain research opportunities • Database of qualitative studies using quantitative datasets used along them to compare experiences of different regions in Australia. • Three data environments: development, staging, production. • Q: how do we deal with reproducibility across disciplines? • The use of data hubs • protocols to bring in other data sets • Workign with the ABS to ensure that multiple Census rounds can be used with a single dataset in PLIDA to be able to capture some family characteristics at different points of people's life trajectories • Feasibility of integrating of QILT surveys (SES, GOS) with TCSI/PLIDA should be assessed • Data on family structures so that individuals from the same family can be linked. • Alternatively, individual-level data linked to some family characteristics. • Critical here will be engagement with data custodian for building trust, enabling access, and encouraging future improvements for data collection (+1) • appropriate training to introduce researchers in one field to methods, data, approaches in a different field (+1) • the need for training • the need for a suite of tools - e.g. GeoSocial - to enable efficient and robust linkage procedure to be executed by researchers • Developing interactive dashboards/tools to map spatial data overlaid onto survey data would offer potential benefits from the perspective of informing policies (e.g. place-based appraoches)

SUGGESTED DIRECTION	RESPONSES
	• tools that offer data curation at a range of spatial and temporal scales - flexible geographies • Providing sustainable access to well-maintained tools • CRUD permissioning in a hierarchy within the central team and the data custodians • Ability to conduct linkages with pre-determined permissions from custodians - Grand Custodian - utopia • Automatic validation of data and metadata. Tools: DBT, Great Expectations, etc. • Organisations to collect data in a useable form (or at all) • common metadata standards across different domains
Connecting data over the lifecourse (for instance by connecting existing data sources about different groups of individuals to give a picture of life course trajectories)	• Much focus is on government data. For this to be successful in the long run, mechanisms are needed to enable linking of different data sets, of verification of the data, as well as to enable creation of new data (field experiments, surveys, etc.) that can be added (+1) • There are many ways to connect - e.g. person, household, cohort, group, community, etc. level. Not all data sets will be linkable at an individual . Thus to connect data over the life course will involve thinking about the different ways to link information across data set (+1) • there is value looking at life course in a broad sense: to include people, families and communities, but also places, land uses and assets such as private vehicle fleets. This would entail significant data linkage to establish the rich yet highly fragmented data landscape in this space - difficult already at the State level but where we really need to achieve this at the national scale. • Full educational histories • Map of relevant datasets that could be linked to provide a national picture of life-course - across disciplines • Designing a longitudinal study of ageaing (or health and ageaing) that would offer some comparabiklity with the US, UK and EU datasets would be particularly beneficial • place-based and other spatially-informed studies of inequality, mobility, and priority areas like regional economic development,

SUGGESTED DIRECTION	RESPONSES
	responses to climate change, community resileience and preparedness. • Datasets matching late life Adult Education populations with early school leavers and broken education participation would help in international comparisons • survye databases from LIEF grants would be good to bring in future • resources - standards,, classifications, vocabularies, code libraries, training to support cross-domain research - e.g. across public health and social science • effective computing environments that can efficiently handle big and complex data
Considering Indigenous data governance in the context of social science data (working out how to incorporate Indigenous data governance principles into integrated public sector administrative data)	• support for Indigneous communities to determine the development and use of data asssets about them. • Ethics for researchers • Record of all data provenance and consent-to-use trail • On-country practices built into project management for joint First Nations and non-indigenous research teams • FAIR & CARE principles • Recognition of ceremony and Indigenous cultural protocols ahead of western thiinking • Indigenous led measures of communtity strengths/assets, wellbeing (+1) • Refer to https://www.pc.gov.au/inquiries/completed/closing-the-gap-review/report/closing-the-gap-review-supporting-paper.pdf • Establishing Australian specific protocols (e.g. Canada has OPAC) • "Establishing australian-specific protocols" - Australia has Maiam Nayri Wingara https://www.maiamnayriwingara.org/ . See also this discussion paper: https://www.lowitja.org.au/news/taking-control-of-our-data-a-discussion-paper-on-indigenous-data-governance-for-aboriginal-and-torres-strait-islander-people-and-communities/ • researcher training on Indigenous governance research jpractice and principles and whey they matter

SUGGESTED DIRECTION	RESPONSES
	• appropriate training for non-Indigenous researchers, advocacy by non-Iindigenous researchers • culturally informed research training for Aboriginal and Torres Strait Islander researchers • Building Indigenous capacity for data governance with elders and community • Data rights and ownership/contextual data access • creation of community-driven standards and classifications. for instance supplement ASGS with Indigenous community and language boundaries

Activity 3: Connections and gaps

QUESTION	RESPONSES
What are the relationships and overlaps between the possibilities we've discussed today?	• Spatial data could be used to augment a broad range of stasets - inclduing longitudinal and cross-sectional surveys in various domains, as well as individual-level admin data (which ofen include some type of information about location) • Building something for Australia could mean starting at the bottom to understand the opportunities and challenges to overcome but with an eye to a national solution that can cater to research taken at differnt levels: e.g. local, state, and national elvel • an international connection .. building something for Australia may also involve building something that is recognized as the gold standard internationally • ABS has been identified as key connector but what about ONDC, other government departments? • "ABS has been identified as key connector but what about ONDC, other government departments?" - Agree, the wall between AIHW and ABS assets is a challenge • There are several emerging themes particular increasing the FAIR-ness of data, increasing capability and collaboration, decreasing duplication and distractions to enable better research outcomes

QUESTION	RESPONSES
	• Building up researcher capability across the social sciences rather than it being limited to typical types of studies, topics, or fields • Use a collaboration process such as https://scccp.net/resources/cobeo-tools/ to explore opportunities and challenges for cross-cultural collaboration before technical concept development & design processes are initiated. • the opportunity to form new academic collaborations through shared research infrastructure • Getting to know social sciences experts who are also interested in health - such cross-sector community would be invaluable • Wrapping a community of practice around the analysis of administrative data for research purposes (forum for discussing issues, indexing publications to get ideas about what is possible, a place to identify those with existing expertise) • Possibly a technology that allow data custodians the visibility on their data and how it is used across different research projects • Identify mid and low level data analysts in large organisations / government departments, and perhaps organise hack fests with them: they understand the idiosyncrasies and opportunities for data aggregation. • Data custodian connections and relationships are critical . Ensureing trust and an ability to minimize the risk of misues or unauthorized access (+1) • Considering what requirements are needed for Data Custodians to feel confident in your ability to mange the data respectful and securely (i.e becoming Accredited under the DATA Scheme) • connecting with policy stakeholders and service providers to understand the value of better use of data and research infrastructure • Connecting sources and custodians of administrative data across disciplines for ease of data access (+1) • building long term reciprocal relationships between data custodians and the research community - ensuring a broadly symmetrical relationship where the data provided to researchers and the products from the research is fed back to the custodian in some form.

QUESTION	RESPONSES
Are there any opportunities we haven't discussed so far in today's session?	• What are the real world challenges we are trying to help address • Job security for academic methodologists (e.g., statisticians) is an ongoing challenge • Regarding health, AIHW have much data access and reporting, have not seen them mentioned explicitly. • Cultural training and engagement • the assumption is that collaboration is the way to go. Think about the sequencing of DNA .. it got done quickly because of a competititve approach. Why not consider a range of projects designed with similar goals but different approaches to enable the most innovation? • Several facilities exist to address many of the researcher requirements mentioned here - how can we work together in a more coordinated way to collectively deliver better infrastructure and outcomes for researchers • a mapping of common aims between researcher and nonresearcher partners in this activtity • Potential public backlash about privacy • Potential public backlash about identification if datasets are linked • issues with linking cohort and admin data - ethics and governance requirments • secure data access that protects data rights and easy to use by research community • types of suitable software/computing environments • It's important (although difficult) to have a 'versioning' system for data and metadata. Examples: Zenodo. • If there's high-frequency data, the solution needs a 'pipeline management system' to update things automatically. Examples: AirFlow, Prefect • harmonising requirements/standards across jurisdictional boundaries (+1) • Mechanism for knowing who else has asked a data custodian for access to data (making it more likely custodians will share, and wheel will not be reinvented) • Difficulty of accessing Commonwealth and state govt data if researcher not working on project for govt.

QUESTION	RESPONSES
	• The details around how to overcome data access barriers, is it even possible in some cases? • The quality and definitions of the same data varies (e.g. AIHW and ROGS report the same variables but because of different definitions the numbers are different) • Start by establishing one big dataset that sets the gold standard for metadata, access, quality - then bring others up to that standard • pays off to have a central team responsible for maintaining metadata and enforcing the gold standard • It is not only about more data but how data is collected that has impact on research questions • automatic data quality validation • Thinking about linkages to other ARDC focus areas - e.g. mediated data; how can we conenct some of the data that could be harnessed via social media etc with other types of data; similarly for transaction-type of data; linkages at a (small-level) spatial unit might be one way to go about it • Linkage between social science data sets and non-social science data sets, as often required to answer questions regarding sustainability challenges e.g., impacts of changes in climate and extreme weather conditions, impacts of changes in major industries (e.g., energy transition/net zero transition) both locally, regionally and nationally. (+1) • Going beyond Commonwealth and State govt data, to industry, non-profits and completing data picture through what they hold • working between different disciplines and research methods to use data in most efficient way • Harmony between research questions and data is what we all want... but there's no way to know if the data our centre holds could be useful to someone else. Some national appraisal of the Big Questions in various HASS fields might be useful. We tend to know our own discipline well and others not at all... • innovation requires not just doing more of what we are already doing but thinking of the next horizon. e.g. other data sources, other mechanisms being introduced to create data, consdiering the ability to create representative synthetic data

QUESTION	RESPONSES
	• Identification of (?smaller) datasets which are suitable for the development of new analytical technologies. For example, the large language models have been developed overseas and need to be localised for productive use. Who holds text data which could localise the various LLMs with research relevance? (BERT / BioBERT) • Consider ecosystems of different kinds of data, with ability to harness multiple kinds of traditional and new (AI) techniques (+1) • the essential value of strong, unified data custodian <> research relationships | that embed trust and a reciprocal relationship where research outcomes are fed back to data custodians • the data custodian relationshp • it seems that creating a linked data asset that is de-identified depends on trust between the data services - currently missing • Clear articulation of value propositions for non-research data holders • Building trust from data custodians involves researchers reassuring them about how data will be used. • A way for researchers to share pro forma documents/discussion topics they had with data custodians to help others not reinvent wheel when approaching a new data custodian • A focus on what data custodians get from this process - how does it give back to them. How do we articulate this to them? • "it seems that creating a linked data asset that is de-identified depends on trust between the data services - currently missing" - not sure if a third party data linker with bi-lateral trust to data holders, could be made to work, might be worth finding some asset owners willing to try