Mr Ryan Winn
Director – National Research Infrastructure Policy
Research and Economic Group
Australian Government Department of Education and Training
ResearchInfrastructure@education.gov.au

10 September 2018

Response to National Research Infrastructure: Humanities, Arts, and Social Sciences Scoping Study

Dear Ryan,

Thank you for inviting the Academy of the Social Sciences in Australia (ASSA) to provide further input into the HASS Scoping Study. ASSA commends the Department of Education and Training for undertaking this initiative and we look forward to participating in this process.

Please find included in this document a response to the Department's request for:

- Themes for scoping the future national research infrastructure requirements.
- Advice on current (or forecasted) researcher infrastructure needs.

The recommendations provided herein are in line with the Department's directive that investment should support leading-edge research and innovation and be accessible to publicly and privately funded users domestically and internationally.

Furthermore, we have included recommendations and a draft Timeline for what we propose to be a Social Science Scoping Study that would form a part of the broader HASS Scoping Study (Item 7).

Rather than repeat background information provided in previous submissions to the Department, this document seeks to provide the recommendations you have requested in a succinct form. A list of relevant previous submissions is detailed at Item 9.

ASSA is at your disposal.

Regards,

Professor Glenn Withers AO
Academy President
Academy of the Social Sciences in Australia
26 Balmain Crescent, Acton ACT 2601

ACADEMY OF
THE SOCIAL SCIENCES
IN AUSTRALIA

Table of Contents

Authors

- Professor Glenn Withers AO – President, ASSA
- Professor Mark Western - Director of Institute for Social Science Research, University of Queensland
- Dr Steven McEachern - Director, Australian Data Archive, Australian National University
- Dr John Beaton – Executive Director, ASSA
- Ms Liz West – Deputy Director, ASSA

1. Summary of Recommendations

A. The development of a stand-alone Social Science scoping study, as a component of the broader HASS Scoping Study, which recognizes the interoperability of social science data and the achievements and potential of systems such as: the Australian Data Archive (ANU); Commonwealth government generalized survey and administrative data investments as described in the Government's 2013 *Essential Statistical Assets* and 2015 *Public Data Policy Statement*; and specialized data assets such as the longitudinal datasets funded by the Department of Social Services and managed by the National Longitudinal Data Centre.

NATIONAL RESEARCH INFRASTRUCTURE THEMES

B. An integrated data management environment, having the following attributes:
   i. Data storage facilities (such as National Computational Infrastructure and Research Data Services) providing high availability and large-volume data storage and back-up facilities.
   ii. Secure access facilities (such as the Population Health Research Network and the ABS Microdata Laboratory).
   iii. Data linkage (through Statistical Data Integration Authorities), and data integration and interrogation environments (such as the Australian Urban Research Infrastructure Network).
   iv. Accredited Data Authorities which build on the responsibilities of integrating authorities to provide advice and technical services, such as data cleaning, de-identification, linkage and safe release.
   v. Data discovery services (such as Australian National Data Service) to enable data discovery through shared platforms and services such as Digital Object Identifiers.
   vi. Sufficient data protection and security through the Five-Safes Framework

C. A stable long-term preservation and curation environment.

D. Effective access for bona-fide researchers, including:
   i. capabilities for open data download (data.gov.au) and mediated data access (Australian Data Archive)
   ii. Availability of secure access environments (such as PHRN and the ABS Data Lab)
   iii. Machine-to-machine data access.
   iv. National protocols for data sharing, access and utilisation such as the new Data Sharing and Release Act.

E. Soft infrastructure in support of innovation in research methods, such as the use of probability-based sampling methods for online surveys, experimental methods for

behavioural economics, and data science methods for description, classification and prediction to enable new forms of research data collection and analysis to be established and supported.

F. Capacity building to promote knowledge of legislative frameworks, methodologies, datasets and technology and infrastructure use throughout the sector.

G. Adequate training for researchers and data providers in the new platform including provision of expert archival assistance to facilitate access.

2. Overview

The ability of Australian social scientists to provide knowledge-based advice to government and effective research outcomes is often dependent upon their capacity to study long term trends and developments and increasingly to predict patterns and outcomes at large and small social and spatial scales. Those studies depend critically upon access to high quality, representative data to describe, compare and predict the social, political, economic, and environmental characteristics of Australia and Australians.

The types of data used in Australian social science research include fundamentally crucial basic data, such as:

- The Australian population census conducted by the Australian Bureau of Statistics (ABS).
- Longitudinal studies of individuals, such as the Household Income and Labour Dynamics in Australia, Longitudinal Study of Australian Children, and Australian Study on Women's Health and on Men's Health, and the Longitudinal Surveys of Australian Youth (funded by various Australian Government departments).
- Nationally representative cross-sectional surveys such as the Australian Election Study and Australian Survey of Social Attitudes, and periodic cross-sectional surveys by the ABS.
- Time series of economic and social phenomena such as unemployment, labour force participation and other statistics produced from the ABS Labour Force Survey.
- Integrated Administrative and other government data collections such as the Multi-Agency Data Integration Project combining data from the Australian Bureau of Statistics, Australian Taxation Office, and Australian Departments of Health, Human Services, Social Services and Education and Training, the Australian Census Longitudinal Dataset and the Mortality data register from the Australian Institute of Health and Welfare, the Business Longitudinal Administrative Data Environment (ABS, Department of Industry, Innovation and Science, ATO, IP Australia) and open data sources such as data.gov.au.
- Business data collections such as the historical records of firms and other organisations held in locations such as the Butlin Archives and recent data from business activity and records.
- Qualitative data collections such as the sub-studies from the Australian Longitudinal Study on Women's Health.
- Linked data such as that provided through the Population Health Research Network and Australian Urban Research Infrastructure Network, and the Australian Census Longitudinal Dataset.

These major forms of traditional data collection can now be complemented by incidental data from social media sources such as Twitter and Facebook and from business analysis of transactions conducted.

Small and mid-sized institutional users of data are finding free-access becoming limited in supply and the commodification of big data and deep analytics tools being affordable only by very large corporations and governments. This will only heighten the need for Australian research infrastructure to accommodate the needs of academic institutions and their researchers.

Crucial infrastructure to support Australian data collection has been established through existing programs of government agencies such as the Australian Bureau of Statistics and Department of Social Services, and funding programs such as ARC, NHMRC and NCRIS. The Australian Data Archive (ADA) also provides access to some of the studies listed above. There are however still many gaps and deficiencies. Importantly, we are concerned that each of these sources is subject to significant risks, including inconsistent funding and even data loss. Furthermore, a significant proportion of Australian data cannot be made available through an open data environment such as data.gov.au at the unit record level needed for much academic and policy research due to privacy and confidentiality requirements.

3. Current Best Practice

Australian Government data is a “strategic national resource that holds considerable value for growing the economy, improving service delivery and transforming policy outcomes for all Australians” (PM&C 2018). Realising this value require researchers and others to be able to work effectively with Commonwealth data to develop, implement and evaluate evidence based policy and practice. This facility in turn requires best practice arrangements for Commonwealth Data Sharing and Release.

Australia is on the path to best practice informed by the Prime Minister’s 2015 Australian Government *Public Data Policy Statement and the Productivity Commission’s Data Availability and Use Inquiry* (2016-2017) which recommended national reforms to overcome barriers and issues with Australia's current data system. In response the Australian Government committed to: a) Establishing a National Data Commissioner to implement and oversee a simpler, more efficient data sharing and release framework; b) Introducing legislation to improve the sharing, use and reuse of public sector data while maintaining strong security and privacy protections; c) Introducing a Consumer Data Right (CDR) to allow consumers to share their transaction, usage and product data with service competitors and comparison services.

The National Data Commissioner (NDC) will oversee the data system within the new data sharing and release (DS&R) arrangements. The DS&R bill to be introduced in 2019 will apply consistent safeguards to Commonwealth DS&R based on the Five-Safes disclosure risk management framework and will create accredited bodies with different roles, skills and expertise, including data custodians, Accredited Data Authorities, and trusted users. The NDC will liaise with the Office of the Australian Information Commissioner (OAIC0 and be advised by a new National Data Advisory Council.

The new DS&R Bill will:

> apply to all Commonwealth entities and Commonwealth companies (as defined under the *Public Governance, Performance and Accountability Act 2013*) and encompass all data collected by these government bodies for any purposes, including government administration, service delivery and research. Data collected from individuals, businesses and other entities, and data generated internally by Australian Government bodies is in scope. There will be appropriate exceptions for national security and law enforcement data. Existing contractual obligations, including around purchased datasets, will continue to apply.

> The Bill will accredit trusted users and data authorities to streamline future interactions. As recommended by the Productivity Commission, the Bill could also enable sharing and release of Australian Government data with States and territories and non-government sectors for specified purposes and subject to certain conditions. (PM&C 2018)

The purpose test for the bill specifically includes research and development with public benefits by research institutions and academics.

The bill will adopt the internationally recognized Five-Safes risk management approach currently used by the ABS, the UK, New Zealand and in Europe, which dovetails with the Privacy Act and the Protective Security Policy Framework for classified government information. It will also complement existing state frameworks such as the South Australian Public Sector (Data Sharing) Act and the corresponding New South Wales. Once the Bill is passed, Australia will have, for the first time, a legislative framework within which all Commonwealth data can potentially be integrated and shared for research, together with a scaleable and efficient framework for increasing data access.

The infrastructure and capacity building to realise value from the new DS&R legislation and current Commonwealth Integrated Data Assets such as MADIP and Blade are core planks in the HASS Social Science infrastructure capability. At the time of writing however, such infrastructure has not yet been established. There exists therefore an opportunity to align the requirements flowing from the establishment of the *Data Sharing and Release Act* with the scoping of the HASS Platform requirements – as there is likely to be significant overlap between the infrastructure requirements of the two programs. In order to get full value from the new legislation, investment needs to be made in Accredited Data Authorities who will provide critical services to safely unlock government data for the research sector. These services need to be underpinned by infrastructure which enables data to be curated, linked and safely shared with researchers.

4. Future National Research Infrastructure

In order to maximise the contributions of the social sciences and others, any investment in national data infrastructure should ensure the following:

1. An integrated data management environment, having the following attributes:
    a. Data storage facilities (such as National Computational Infrastructure and Research Data Services) providing high availability and large-volume data storage and back-up facilities.
    b. Secure access facilities (such as the Population Health Research Network and the ABS Microdata Laboratory).
    c. Data linkage (through Statistical Data Integration Authorities), and data integration and interrogation environments (such as the Australian Urban Research Infrastructure Network).
    d. Accredited Data Authorities which build on the responsibilities of integrating authorities to provide advice and technical services, such as data cleaning, de-identification, linkage and safe release.
    e. Data discovery services (such as Australian National Data Service) to enable data discovery through shared platforms and services such as Digital Object Identifiers.
    f. Sufficient data protection and security through the Five-Safes Framework
2. A stable long-term preservation and curation environment (the Australian Data Archive).
3. Effective access for bona-fide researchers, including capabilities for open data download, secure access, and machine-to-machine data access.
4. Soft infrastructure in support of innovation in research methods, such as the use of probability-based sampling methods for online surveys, experimental methods for behavioural economics, and data science methods for description, classification and prediction to enable new forms of research data collection and analysis to be established and supported.
5. National protocols for data sharing, access and utilisation such as the new Data Sharing and Release Act.
6. Capacity building to promote knowledge of legislative frameworks, methodologies, datasets and technology and infrastructure use throughout the sector.
7. Adequate training for researchers and data providers in the new platform including provision of expert archival assistance to facilitate access.

Such a system, through extensive linkages, would support the collection, analysis, curation, and utilisation of Australian data. It would bridge all data users and providers including academics, government, business, industry, NFPs, community organisations and even individuals.

There is also compelling and growing demand in higher education institutions worldwide to provide education and training in New Analytics and Citizen Analytics. These courses will help to serve the needs of industry and business to have work-ready graduates who are informed and capable in areas of data analysis appropriate to massive data sets and emerging technologies.

Why is this important?

Institutions including those of industry, business, government, and community organisations all require a sophisticated understanding of data relating to their constituent parts and the populations they serve. There are a number of reasons as to why national effective and sustained national funding of research infrastructure is crucial for all data users. In addition:

1. There are legal requirements across all sectors for the secure storage of data relating to their activities.
2. Research results across academic sectors require effective curation, as do institutions and individuals on public funding, including for purposes of replication and accountability.
3. Regulatory agencies require substantiated records of research activities, methods, and outcomes.
4. The maintenance of an integrated system of data management is a highly efficient means of providing generations of researchers with comparative evidence to assist them in the production of sophisticated and useful analysis.
5. An integrated national data infrastructure system can maximise data security mechanisms and regulate effective access.
6. Australia will do well in its international partnerships to provide a best-practice data management environment where trust is an imperative.

> New Zealand for instance, has developed its Integrated Data Infrastructure, a research data environment containing microdata about people and households from Statistics NZ surveys Census, and non-government organisations. The IDI holds over 166 billion pieces of information, taking up 1.22 terabytes and is continually growing. Researchers access the IDI through the NZ Statistics IDI Data Lab or can apply to set up a secure IDI lab in their own organisation. IDI access is governed under the Five-Safes Framework, covered under relevant legislation, and also subject to an overarching and periodic Privacy Impact Assessments and New Zealand's more general Data and Information Governance Strategy. New Zealand is currently ranked 4th in the world on the Open Data Barometer, ahead of

Australia at 10, and is a member of the Open Data Institute Leaders network for government open data initiatives around the world. Another example of an integrated data system is the Survey of Health, Ageing and Retirement in Europe (SHARE). SHARE is harmonised with the U.S. Health and Retirement Study and the English Longitudinal Study of Ageing and has become a role model for several ageing surveys worldwide.

5. Shovel-Ready Initiatives

Australia has many of the building blocks in place for a national integrated data infrastructure for structured administrative and survey data. We have the national data investments in the Government data assets, through the ABS, Data Integration Partnerships for Australia, and specialized longitudinal data assets. We also have the beginnings of the Legislative framework through the 2019 planned Data Sharing and Release legislation, and governance through the Office of the National Data Commissioner, infrastructure for access through the ABS Data Lab and virtual Data Lab.

Through the Australian Data Archive (ADA) we have a Core Trust Seal certified repository, based in the ANU Centre for Social Research and Methods (CSRM) at the Australian National University (ANU), with over 6000 datasets from more than 1500 projects and studies from 1838 through until the present day. The ADA was established with a brief to provide a national service for the collection and preservation of digital data relating to social, political and economic affairs and to make these data available for further analysis.

We also have the Australian Urban Research Infrastructure Network (AURIN), the NCRIS-funded facility, for enabling the spatial analysis and visualisation of data from urban research programs (which includes some social science disciplines). NCRIS investments in AURIN, along with ANDS and PHRN, have resulted in high quality infrastructure for data discovery services, data linkage and secure access facilities for the social sciences. The needs for High Performance Computing in the social sciences are different from other domains, but can likely be accommodated by existing NCRIS HPC infrastructure (with the possible exception of the analysis of data generated from social media activity and the Internet of Things).

What we don’t have is an integrated national data infrastructure that joins these capabilities, has ongoing arrangements for data curation and storage, and is searchable and accessible for researcher via modes of interaction and analysis that the community wants and that facilitate research data management, reproducibility and collaboration (within and between universities, agencies, departments and internationally).

An integrated research infrastructure for the Social Sciences would provide:

- The capacity to find, access, preserve, document and disseminate quantitative and qualitative data sources such as surveys, in-depth interviews or the results from experimental trials, in a stable, long-term curation environment;

- Systems and tools for the capture and analysis of real or near real time data sources, including social media, websites and web traffic, and the Internet of Things;
- A cost-effective and accessible data integration and linkage environment;
- Secure data facilities (physical and virtual) for enabling access to sensitive data, such as linked survey and administrative data produced in the above data linkage environment;
- Research data infrastructure to allow high quality nationally representative surveys of the Australian population using an academic probability online panel (for example modelled after the GESIS Panel research data infrastructure for online data collection in Germany);
- Question and variable banks that allow for: (i) creation of new datasets from existing data; and (ii) extraction of questions for use in new data collections; and
- Infrastructure in support of innovation in research methods, such as the use of probability-based sampling methods for online surveys, and experimental methods for behavioural economics.

Such a platform would also link and integrate with the key national research infrastructure, including:

- Data discovery services (Australian National Data Service) to enable data discovery through shared platforms and services such as DOIs;
- Data storage facilities (National Computational Infrastructure, Pawsey and Research Data Services) providing high availability and large-volume data storage and back-up facilities;
- Secure high capacity networks (AARNet) for transfer of data between researchers and facilities, and to enable remote access to secure facilities;
- Data linkage through Accredited Integrating Authorities (e.g. ABS, AIHW, AIFS), and data integration and interrogation environments (such as the PHRN and AURIN); and
- Secure access facilities (e.g. PHRN, ABS, AURIN) for analysis of data in secure, protected virtual environments.

In the short-term, investment should focus on increasing the value of the existing institutional investments being made in social science data infrastructure and filling gaps, such as:

- Substantially increase the social science data archiving capacity of the Australian Data Archive, including the development of a comprehensive collection of high

value qualitative data which is made available for re-use and the establishment of a question and variable bank;

- Improved Integration of existing social science research infrastructure – particularly ADA, AURIN, Atlas of Living Australia, ANDS/ARDC, Analysis and Policy Observatory (APO);
- The development of the research data infrastructure for a rigorous nationally representative probability online panel, hosting major national and international social surveys. The panel would be linked to the data archive including the questionnaire and variable bank;
- Establishment of infrastructure for innovation in research methods; and
- Extension of the existing secure access facilities (PHRN/AURIN/ABS/ADA) to enable improved data linkage, integration and access to sensitive government administrative and other data sources (e.g. health records, biomedical data).

6. The Case for a Social Science Scoping Study

Through this consultation process, two labels have been frequently applied to the HASS sector: broad and disparate. While the HASS sector is no broader nor more disparate than the STEM sector, it is also not homogenous. The data structure and uses for the social sciences differ greatly from those in the Arts, Humanities, or GLAM institutions, and come with their own challenges and opportunities. For this reason, and given the successes and opportunities detailed above, ASSA recommends the creation of three sub-studies that would feed into the main Scoping Study: 1) humanities, 2) arts, 3) social sciences.

Importantly, the Department may wish to consider a data platform (digital data infrastructure) that treats social science and arts/humanities data separately. This platform should allow interoperability between the modules of data.

7. Development Process

The Social Science Scoping Study proposed in this document could serve as a model for both the other sub-studies and the broader HASS study. ASSA proposes the following timeline for undertaking this element.

	Phase	Detail	Estimated Duration
1	Preparation	• Identification of Expert Working Group (EWG) and Advisory Panel (AP). • Identification of stakeholders for consultation and nomination of representatives. • Development of project budget and final timeline.	1 month
2	Audit	• Audit of existing research infrastructure. • Identification of data providers and users. • Identification of overlap, interoperability, global best practice, access and knowledge gaps.	3 months
3	Consultation	• Bringing together (workshop) representatives from stakeholder institutions, data providers/users, information technology designers to identify priorities for capabilities of new research infrastructure. • Agreement of needs, priorities, and assumptions. • Identification and scoping of governance model and data management resource needs.	1 day workshop 1 month preparation
4	Development	• Information technology development of data platform (or enhancement of existing platform) in consultation with EWG. • Departmental development of governance structure in consultation with EWG and AP.	9 - 12 months
4	Testing	• Testing by data subject matter experts (data providers and users).	
5	Launch		

8. Stakeholder institutions

- Academy of the Social Sciences in Australia (ASSA)
- Australian Data Archive (ADA)
- Institute for Social Science Research (ISSR)
- Centre for Social Research and Methods (CSRM)
- Australian Bureau of Statistics (ABS)
- The Office of the National Data Commissioner (ONDC)
- Australian Urban Research Infrastructure Network (AURIN)
- The Melbourne Institute
- Australian National Data Service (ANDS)
- Australian Research Data Commons (ARDC)
- Population Health Research Network (PHRN)
- Atlas of Living Australia
- Centre of Policy Studies, Victoria University
- Analysis and Policy Observatory (APO)
- Commonwealth Scientific and Industrial Research Organisation (CSIRO)
- Australian Institute of Health and Welfare (AIHW)
- Australian Institute of Family Studies (AIFS)
- Centre of Excellence in Population Ageing Research (CEPAR)
- Centre for Aboriginal Economic Policy Research (CAEPR)
- The Grattan Institute
- The Australian Research Council Centre of Excellence for Children and Families Over the Life Course
- Department of Social Services
- The Social Policy Research Centre
- Telethon Kids Institute

9. Related Submissions

For background information, please consult the following submissions and publications:

- ASSA and ADA/CSRM Submissions to 2016 National Research Infrastructure Roadmap Issues Paper (https://submissions.education.gov.au/Forms/National-Research-Infrastructure-Capability-Issues-Paper-Submissions)

- ASSA and ADA/CSRM submissions to NCRIS Roadmap Draft (https://submissions.education.gov.au/forms/2016-strategic-roadmap/pages/index)

- ADA submissions to the Productivity Commission Inquiry into Data Availability and Use (http://www.pc.gov.au/inquiries/completed/data-access/submissions)

The Academy is available at any time to further discuss this submission.

Academy of the Social Sciences in Australia
26 Balmain Crescent, Acton ACT 2601
GPO Box 1956, Canberra ACT 2601
P: +61 2 6249 1788
John.Beaton@assa.edu.au
Liz.west@assa.edu.au