Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Dr Steven McEachern and Professor Matthew Gray**
Title/role	**Director, Australian Data Archive (SM)** **Director, ANU Centre for Social Research and Methods (MG)**
Organisation	**Australian Data Archive, ANU Centre for Social Research and Methods, The Australian National University**

Australian Data Archive summary of key issues

The Australian Data Archive was established in 1981 as the national social sciences data archive. The archive currently preserves, documents and makes available for re-use about 6,000 social science data sets. The NRIC Issues Paper identifies the Australian Data Archive as a national infrastructure.

While the Australian Data Archive goes some way towards providing a national social science data infrastructure, Australia's current infrastructure is inadequate for the contemporary data environment and we are slipping behind developments in the UK, US and a number of continental European countries. A significant investment in Australia's national social science data infrastructure is urgently required.

Access to cross-sectional and longitudinal survey data, administrative by-product data and incidental data from social media and other sources such as GPS data from connected devices is essential to social science research. While there has been an enormous expansion in the amount of data available, much of this data is difficult to access or is not available to the broader research community.

All of the key social science data sources are subject to risks resulting from inconsistent funding and data loss. Many of the existing data collections are funded on a short-term basis. Changes in data collection and analysis also require innovation in research methods to ensure they are consistent with the ongoing needs of Australian social research and public policy.

What is needed?

While significant public investment has been made in the generation of social science data, provision of funding for a national social science data infrastructure would support the core needs of the social science community in Australia by providing:

- The capacity to **find, access, preserve, document and disseminate** more of the surveys currently being undertaken in Australia;
- A **stable curation environment** for the long-term preservation and storage of social science data;
- Dedicated **data dissemination and access environment** for social science data, including capabilities for open data download, secure access and machine-to-machine data access

from the Australian Data Archive to various access facilities and systems for data analysis (for example, AURIN, PHRN);

- A **data integration environment** to enable social science data to be enhanced through linkage of data, particularly survey and administrative data sources, but also inclusion of spatial characteristics;
- **Secure data facilities** (physical and virtual) for access to highly sensitive data such as linked survey and administrative data;
- Infrastructure to support **innovation in research methods**, such as the use of probability-based sampling methods for online surveys (e.g. the GESIS Panel for online data collection in Germany[1]), and experimental methods for behavioural economics, to enable new forms of research data collection and analysis to be established and supported; and
- **Questionnaire 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 (e.g. to build a new questionnaire) in an efficient and standardised way

These services can be enhanced through the **integration** with existing infrastructures:

- **Data storage facilities** (National Computational Infrastructure and Research Data Services) providing high availability and large-volume data storage and back-up facilities;
- **Secure access facilities** (the Population Health Research Network and the Australian Bureau of Statistics (ABS) Microdata access facilities[2]);
- **Data linkage** (through Accredited Integrating Authorities[3]), and data integration and interrogation environments (such as the PHRN and Australian Urban Research Infrastructure Network);
- **Secure high capacity networks** (AARNet) for transfer of data between researchers and facilities, and to enable remote access to secure facilities; and
- **Data discovery services** (Australian National Data Service) to enable data discovery through shared platforms and services such as DOIs.

Why is this important?

There are a number of reasons as to why national funding of this research infrastructure is crucial.

- The storage of data is a requirement of ethics committees and at present it is a challenge for both individuals and institutions to provide the necessary data repositories;
- It is important that data be available for the purposes of replication and extension of analysis in published papers. This is increasingly being recognised in the physical sciences (e.g. 'Climategate') and is increasingly being recognised as an issue for the social sciences. This is vital to scientific credibility;
- Opens up data for secondary analysis, particularly for PhD students and early career researchers;

1 http://www.gesis.org/en/services/data-collection/gesis-panel/

2 http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Microdata+Entry+Page

3 https://statistical-data-integration.govspace.gov.au/roles-and-responsibilities/integrating-authorities/

- It is costly to collect high quality survey data and so it is important to avoid duplication and maximise the value of data that is collected. There are costs associated with the data collection and in terms of respondent burden;
- Access to ABS data is often highly constrained and so academic collections are critical to research;
- Australia is well behind the UK, US and most of Europe on open data. This is impacting Australia's ability to be competitive and its standing in the HASS discipline; and
- There is a need to encourage a new generation of researchers and public sectors who are committed to sharing data.

Response to consultation questions

Question 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national research infrastructure.

The governance characteristics listed in the Issues Paper appear to be appropriate. A gap in the list of characteristics is shared approaches, such as reference models and policies, where common issues exist across infrastructures and capabilities. An example of this need relates to data access models.

A recent international development among social science data infrastructure providers has been the adoption of trust models to support access to data for research and broader use. A potential model for this is the Five Safes model developed at the UK Data Archive in conjunction with the UK Office of National Statistics[4]. The Five Safes model has the following elements:

1. Safe People: Can the researcher(s) be trusted to use the data in an appropriate manner?
2. Safe Projects: Is the data to be used for an appropriate purpose?
3. Safe Settings: Does the access environment prevent unauthorised use?
4. Safe Data: Is there a disclosure risk in the data itself?
5. Safe Output: Are the statistical results non-disclosive?

This model has recently been adopted by both the Australian Bureau of Statistics (ABS) and Statistics New Zealand as their reference model for assessing new and existing data access methods for their products and services.[5]

The Australian Data Archive has been promoting the adoption of the Five Safes principles as a reference model for data access by government agencies and university and other research groups. Widespread implementation of this type of model would increase trust and confidence in the procedures used for access to sensitive data, and to provide a common reference model for those providing access to such data.

Question 3: Should national research infrastructure investment assist with access to international facilities?

Question 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

Response to Questions 3 and 4.

The Australian Data Archive supports the use of national research infrastructure investment to increase the access of Australian based researchers to international facilities. Access to comparable international data is central to social science research projects and often requires access to archives in other countries.

Access to international facilities does not replace the need for Australian social science data infrastructure for Australian researchers. The existing "international facilities" are in fact not a single

[4] http://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf

[5] http://www.abs.gov.au/AUSSTATS/abs@.nsf/be4aa82cd8cf7f07ca2570d60018da27/e4d483bab4e1ad93ca257f4c00170bb6!OpenDocument ; http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/keep-data-safe.aspx

facility. They are generally a federated or collaborative network, with one or more nodes established in each country to provide the national contribution to an international network.

The three EU ERIC infrastructures in the social sciences (discussed further in our response to Questions 24-26 and 'Other Comments') all have this characteristic – an international hub for coordination of facilities and activities, with national infrastructure to enable the collection, management and dissemination of data at the national level. Thus there is a need here for both an Australian social science infrastructure, and a mechanism to connect this infrastructure to the broader international community.

The model of country-specific data archives has arisen because social science survey data can be very sensitive and confidence of depositors that confidentiality will be preserved and the data sensitivities be managed appropriately is higher for a national data archive. These are tailored to the needs of the social science community within each country, but with a capacity to engage in international collaborative networks to operate as international facility. Model examples for any Australian infrastructure include the UK Data Service, GESIS Leibniz Institute for the Social Sciences in Germany, and the Norwegian Center for Research Data[6], which all operate as national social science data infrastructures, but are also service providers in the EU-wide CESSDA ERIC infrastructure.

Question 5: Should research workforce skills be considered a research infrastructure issue?

The Australian Data Archive supports the view that research workforce skills are a research infrastructure issue – that a skilled research workforce capable of making use of the infrastructure is one of the necessary conditions for a useful and relevant research infrastructure. In the case of data infrastructure, this is even more apparent. Researchers must have the skills to enable them to access, manage and analyse the data available in a manner suited to the conduct of world-class research within their discipline.

A clear example of this in the social sciences is in the area of big data or observational data (using the terminology of the Issues Paper). The growth in the volume and availability of big data sources continues at rapid rates, and provides huge potential (and significant current value) as a research resource. The availability of such data is however in forms (such as JSON) or accessed through new mechanisms (such as APIs) that are often unfamiliar to researchers within the social sciences. The development of the relevant skills to be able to utilise these new data sources is critical to their adoption, acceptance and use within a research community.

To this end, various social science faculties within Australia and internationally have established both short courses and more detailed degree courses in Data Science, Data Analytics and related fields, often in collaboration with relevant disciplines. An example of this is the ANU's Masters of Applied Data Analytics. The UK Data Service (the UK parallel of the Australian Data Archive) has also established short courses in programming and big data to enable effective use of their big data infrastructure.

---

[6] See: UK Data Service - https://www.ukdataservice.ac.uk/; GESIS - http://www.gesis.org/en/home/, Norwegian Center for Research Data - http://www.nsd.uib.no/nsd/english/index.html

Question 6: How can national research infrastructure assist in training and skills development?

Research infrastructures have a significant role to play in the support of training and skills development for researchers. Data infrastructures such as the Australian Data Archive provide real-world data for use in teaching and training programs, and many social science data archives have developed dedicated resources for teaching modules using research data.

In addition staff based within social science research infrastructure facilities often have developed specific expertise in their infrastructure that can and should form part of research user training. For example, staff from the Australian Data Archive, provide specialist teaching in areas such as data management and archiving, and data storage and security.

Finally, the development of training programs and workshops is a critical element in the operations of many data infrastructures – in order to make users aware of what data is available and how it can be used.

Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?

In the social sciences context, the technical resource commitments required for access to facilities are relatively small. Datasets are generally in the megabyte to gigabyte range, and computation resources required in most cases are relatively small. While the size of datasets and computational resources required is expected to grow, they will remain relatively small compared to the requirements of the STEM disciplines.

The greater challenge for social science and humanities is data access and integration which requires being able to:

- Find (or collect) data suited to a researcher's research question;
- Access that data in a timely way, with appropriate access and security controls, in an efficient and suitable format; and
- Integrate that data with other datasets in an ethical and scientifically valid way.

These characteristics are consistent with an access model based on the provision of **free access** for usage, rather than a cost-recovery or merit-based model. We would also note however that this does not necessarily suit an **open data** model, given the privacy and confidentiality considerations of many social science studies. These studies are often collected under conditions of anonymity of participants and mediated or restricted access to data as part of the ethics approval for the research project.

Question 11: When should capabilities be expected to address standard and accreditation requirements?

The Australian Data Archive has a long-standing commitment to the need for standards and accreditation requirements for our research infrastructure, and would recommend these form part of any capability's service requirements.

We note particularly the inclusion of the FAIR principles[7] in the Issues Paper (p.48) that data is "Findable; Accessible; Interoperable; and Reusable". Implicit within these principles is the need for

[7] https://www.force11.org/group/fairgroup/fairprinciples

common standards and practices that are shared by members of the discipline to enable each of these principles to be enacted. For example, for data and metadata are "retrievable by their identifier using a standardized communications protocol" (FAIR principle A1, emphasis in original) requires the use of a standard that enables at least readability and actionability by machines or by humans – or ideally both.

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

The Issues Paper (Section 5) identifies the potential value of linking health dataset with social science data sets such as justice, education and geospatial datasets. The Australian Data Archive strongly supports this suggestion because of the enormous potential gains from better linkages between health and social science data.

Maximising the value of linking health and social science data sets requires the data, the data linkage environment and the tools to link the data. As noted in the Issues Paper, the services provided by PHRN, including data linkage and the SURE access environment are important enablers to achieve this value. The third element – data – however includes both health data and ""data sets outside of health such as justice, education or geo-spatial" (section 5.1.1). Investment in national social science data infrastructure will increase the range of "non-health data" available to be linked to health data, increase the quality of data, reduce costs and reduce timeframes for data linkage.

The foundation for enabling access to such research and administrative data currently exists. The data is often available from data archives and agencies such as the Australian Data Archive, and agencies such as the ABS and the Australian Institute of Health and Welfare (AIHW). The data linking environment has been developed through NCRIS infrastructures such as PHRN and AURIN, along with the administrative and other procedures associated with the Human Research Ethics Committees under the NHMRC ethics guidelines[8], and the Federal and State data linkage authorities. The tools are the relevant databases, and statistical and spatial analysis packages available.

It will be important that national infrastructure investment in this area leverages and extends the existing data infrastructure, as well as enabling the use of that data within integration environments such as PHRN and AURIN. For example, the Australian Data Archive is currently working with AURIN to develop a machine-to-machine capability for the delivery of data into the AURIN environment that is managed through Australian Data Archive facilities (hosted on NCI systems).

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

As noted in our response to Question 15, this is a capability area that overlaps with major elements of the Understanding Cultures and Communities capability. The Social Determinants of Health framework developed by the World Health Organisation[9] continues to be the predominant framework for understanding health disparities in Australia and internationally. This requires involvement of the social sciences.

---

8 https://www.nhmrc.gov.au/guidelines-publications/e72

9 http://www.who.int/social_determinants/en/

Similarly, the potential for data and tools and infrastructure developed in the Health and Medical Science capability to be applied in social science fields such as demography, public policy and indigenous studies suggests the need for significant consultation between the two capabilities and potential for shared infrastructure.

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?

Response to Questions 24, 25 and 26:

The Australian Data Archive supports the suggestion in Section 8.3.2 of the Issues Paper for a "community managed cultures and communities research infrastructure capability". As the Issues Paper notes, this can build on the work of ATSIDA and the National Centre for Indigenous Genomics in this area, which we believe provides a solid foundation for the establishment of such a capability.

There is a need for some clarification of the points made in the Issues Paper with regard to the core requirements of the humanities, arts and social sciences (HASS), the disciplines most relevant to the "Understanding Cultures and Communities" (UCC) capability, and the role of Australian Data Archive and ATSIDA within this capability. There are also significant international developments that are worth using as a guide the development of the UCC capability Roadmap.

The Issues Paper (Section 8.3.1) suggests that there is potential for the Australian Data Archive to play a greater role in providing digital humanities capability. While the Australian Data Archive currently does support humanities research (for example, indigenous studies, history), there is potential for the archive, in collaboration with other humanities services (such as PARADISEC) to provide a greater level of support in the use of qualitative data sources including images, text, audio and video.

There is a need to expand the range of research areas within the UCC capability to include a broader range of social science disciplines. Many of the current and desirable capabilities within the UCC capability are focussed primarily on Humanities and Arts disciplines.[10] For example, of the identified current and future needs, only the current capabilities in Urban Settlements (Section 8.2.1) and the desired capability in Social and Behavioural Science Innovation (Section 8.3.3) have a social science focus, and these reflect only a small portion of social science research in areas of behavioural economics and housing and urban research.

[10] The Issues Paper makes no separate reference to the social sciences – combining them either with Humanities and Arts (in various locations), or with the behavioural sciences (Section 8.3.3).

The classification of some content within the Urban Settlements capability may also limit its scope somewhat. For example, observational data is mentioned in “Urban Settlements” – but a significant proportion of this data, such as social media (for example) is not exclusively urban, and in some research contexts it is the disconnect from the geographic which is of interest (such as studies of online communities, virtual environments, political networks and the like).

There is a strong need to establish the requirements in this capability in a way that incorporates the range of needs of specific disciplines and areas. There is the potential that a single capability that seeks to addresses such a broad range of requirements across the HASS disciplines will result in services that are ill-suited to the requirements of any specific disciplinary community within HASS. While there are some common requirements across the HASS disciplines, they also differ in numerous ways, some of which are described in Table 1.

Table 1 highlights that there are some common infrastructure needs that can be aggregated across disciplines, particularly within the social sciences. As noted above, the Australian Data Archive’s core data providers and user communities cover a number of disciplines across social sciences, health and business and economics. Researchers in many social science disciplines share a common use of statistical data sources (such as surveys and censuses, administrative sources and observational data) that have common requirements across data formats, access procedures and research software. Similarly, the data sources used in the analysis of qualitative materials by policy researchers, such as images, audio, text and video are often comparable to that used by humanities researchers in history or cultural studies, and may therefore have shared requirements.

Table 1 Dimensions of research and data needs in HASS disciplines

Dimension	Social Sciences	Humanities and Arts
Data types	Statistical data (spreadsheets and statistical packages) Images, text, audio, video Observation data (social media, smart devices, GPS) Administrative by-product data (government records, store receipts, program data)	Images, text, audio, video
Data access	Mediated and restricted	Open to mediated
Research software	Statistical packages Spatial analysis packages Databases CAQDAS (Qualitative data analysis) packages Data mining Productivity suites (eg., MS Office)	Productivity suites Text mining Natural language processing Databases Visualisation tools
Research Ethics Approvals	Regularly required	Rarely required
Research outputs	Journal articles Books and monographs Reports	Journal articles Books and monographs Visual displays Exhibitions and performances

There are number of underlying services and facilities that can be shared across these communities. These are considered further below in responses to Sections 10 and 11 of the Issues Paper, but include:

- Digitisation facilities, such as image, audio and video digitisation;
- Data processing capabilities, including text, images, audio and video;
- Data storage services;
- Data integration services; and
- Data access facilities.

Each of these services would then require specific tailoring to the needs of the particular research community, potentially with the "digital tools and virtual laboratories" outlined in the Issues Paper (Section 11.2.3).

Underpinning Research Infrastructure

Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?

The Australian Data Archive supports the set of directions and capabilities identified in the Issues Paper as critical to maintenance and future development of Australian research infrastructure within our user communities. The Australian Data Archive in particular is dependent on the high performance computing, high capacity network and access and authentication infrastructures currently provided under NCRIS and related funding. Social science researchers also make significant

use of current geospatial systems and would be high volume users of the proposed digitisation facilities.

One potential development that would significantly benefit the social science community is a stronger national capability around secure infrastructure. This would enable social science research and data management within Australia. The specific infrastructure requirements of the social sciences are detailed elsewhere in responses to Q33 and our Summary Statement at the start of this submission.

Such a capability would also enable improved access to international facilities for Australian social scientists, by enabling links to international data archives and facilities to be built. Two examples here would be highly beneficial for Australian social science research:

- the recent Data Without Boundaries project[11], which established protocols and infrastructure for secure data access for researchers across 11 data archives and 10 statistical agencies throughout the EU, and
- the recent collaborations between Cornell University, University of Michigan and UC Berkeley and the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Germany[12], for access to German labour market data through a remote access environment.

Both of these programs enable remote access by international researchers to data which can only be housed in the country of origin for legal reasons – and are dependent on secure data transfer over international networks.

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

The ongoing use of access and authentication systems, high capacity networks and geospatial systems will continue to be a critical resource for most Australian social science researchers. The Australian Data Archive supports efforts in these areas that enable access to the international social science community for Australian researchers.

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

The Australian Data Archive is supportive of the proposed underlying infrastructures in the current Issues Paper. There however is the need to consider the means (hardware, software, metadata and provenance) required to connect these infrastructures in a transparent and efficient manner.

For example, the availability of the proposed digitisation infrastructure needs to be aligned with:

- Relevant data storage infrastructure for storing the digitised content;
- Secure high capacity networks for transferring the data from the digitisation facilities to the storage; and
- A provenance capture infrastructure for capturing the relevant provenance information associated with the digitisation process (e.g. agents, activities and entities if using the W3C PROV data model).

11 http://www.dwbproject.org/

12 https://ciser.cornell.edu/IAB.shtm

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

The Australian Data Archive strongly supports the proposed model identified in on p.48 of the Issues Paper. In the social science community, we see the Australian Data Archive as providing a key infrastructure in the trusted data and collaboration and dissemination services proposed in the model on p.48.

Continuing from our response to Questions 24 above, some of the core infrastructure needs in the social science community are:

- Data capture and processing: digitisation facilities (e.g. for images, audio and video digitisation, and some physical samples), text processing capabilities;
- Data storage services;
- Data integration services; and
- Data access facilities

We would particularly identify the need for infrastructure services that provides appropriate levels of secure access to data for analysis as a critical requirement in the social sciences community. These may be through multiple arrangements:

- Large confidentialised survey datasets and administrative data sources (and potentially also culturally sensitive qualitative content such as interview recordings) would benefit from an access environment such as SURE, that allows remote analysis of research data in a secure environment consistent with government security standards such as the Australian Signals Directorate
- These data may be aggregated (often on a spatial or demographic basis) in combination with other data, in AURIN and similar environments
- The data may also distributed to researchers directly for simple desktop analysis where there is low risk of disclosure of personal information
- The data may then be linked with other sources to enable the use of additional covariates in modelling

The Five Safes model (see response to Question 2) is informative here in aligning the data of interest with an appropriate “safe setting” for access.

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Continued involvement in international data networks such as the Research Data Alliance is critical to ensuring ongoing development of collaborative cross-disciplinary infrastructure for data management, access and dissemination. There are also international disciplinary collaborations on standards and practices that could be supported under this Roadmap (such as the Data Documentation Initiative in social science, and TEI in several humanities domains).

Australian social science researchers also currently participate in many international data collection networks to enable comparative analysis, such as the World Values Survey, International Social Science Program and Comparative Study of Electoral Systems that require national data collection to

contribute to international programs (in a similar way to IMOS's provision of Australia's contribution to international marine science data collection). Networks for these programs already exist that could be readily enhanced through the provision of research infrastructure or partnering with parallel European and North American programs.

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

The Australian Data Archive support the basic principles outlined in Section 11.2.1 on "Better Managed Research Data", but would recommend a more detailed consideration of the expectations associated with this.

We believe there are significant opportunities in the social sciences for enabling "better managed research data", particularly to support more automated processes. However these opportunities need to take account of both the technical capabilities of the researchers involved and the resources available to them to achieve this. The development of a maturity model for understanding the migration path for researchers, projects and disciplines might provide a pathway forward for understanding the likely use of NRIC capabilities, and how these can be developed over time.

A social sciences example is in the area of comparative social attitudes surveys across countries. There have been significant Australian and international efforts to improve the creation, management and dissemination of this data through its lifecycle, particularly, through programs such as the European Social Survey (discussed further below). These programs have enabled some development of end-to-end data and metadata management and provenance, but still require significant amounts of human curation and management to deliver the final data products to researchers. An understanding of the key technical, administrative and skills requirements involved in this research data management process may then assist in engaging researchers in related projects with NRIC capabilities in the long term.

Other comments

This section provides further information about the current and future social science research data infrastructure needs.

1. International best practice

Australia's social science data research infrastructure is rapidly falling behind that currently operating in the US and Europe with a number of countries having invested in coordinated and well-resourced data infrastructures. This infrastructure is generally provided via coordinated publicly funded national centres. The European national centres are coordinated through EU Framework funding for research infrastructure (known as a European Research Infrastructure Consortium or ERIC) to establish pan-European infrastructures. The Australian social sciences sector can learn a lot from the European experiences and models. There is significant potential for partnering with these institutions to enable both data sharing and access, development of common technology and infrastructure, and for Australia to leverage off the significant investments that have been made.

There are several key European infrastructures that can be considered best practice models:

- Consortium of European Social Science Data Archives (CESSDA): A network of national social science data archives with a common core of technical infrastructure, policy and data and metadata standards
- The European Social Survey (ESS): a shared infrastructure for the conduct of a Europe-wide social attitudes and values survey, with common instrumentation, methods and collection processes
- Survey of Health, Ageing and Retirement in Europe (SHARE): a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of approximately 123,000 individuals (more than 293,000 interviews) from 20 European countries (+Israel) aged 50 or older (http://share-project.org)
- SERISS: A four year program to support integrative activities that support all of the above three ERIC infrastructures with common tools and practices – such as a questionnaire development environment and online survey panel facility

In addition, there are often smaller national or multi-national programs that provide shared infrastructure for researchers across several countries. Examples include:

- CLOSER – the UK program funded by the Economic and Social Research Council to enable shared infrastructure, training, data harmonization, data access and linkage, and discovery for the UK's major longitudinal panel studies;
- International Public Use Microdata Samples – an infrastructure provided by the Minnesota Population Centre to support access to samples of national censuses from 80 countries around the world; and
- The Digital Panopticon – a UK-Australian collaboration using digital technologies to bring together existing and new genealogical, biometric and criminal justice datasets held by different organisations in the UK and Australia, exploring the impact of the different types of penal punishments on the lives of 66,000 people sentenced at The Old Bailey between 1780 and 1875.

2. The need for a national collaborative social science data infrastructure

The core arguments for the support for a social science data infrastructure centre around three basic needs:

- Sound scientific practice;
- Return on public investment in social surveys; and
- Public trust in social science research

2.1 Sound social science research practice

High quality social science research requires that data be available to other researchers in order to test, verify, falsify and replicate empirical studies. This is also important to maintaining public confidence in the findings of research on important social and economic issues. A core social science data infrastructure supports the achievement of this.

The importance of the archiving of data is recognised by the Australian Code for the Responsible Conduct of Research (ACRCR):

> *"The central aim is that sufficient materials and data are retained to justify the outcomes of the research and to defend them if they are challenged. The potential value of the material for further research should also be considered, particularly where the research would be difficult or impossible to repeat." (NHMRC/AVCC, 2007, s.2.1)*
>
> *if the work has community or heritage value, research data should be kept permanently at this stage, preferably within a national collection (NHMRC/AVCC, 2007, s.2.1)*

A national data infrastructure provides a cost effective and secure way of allowing research to meet their obligation to retain and make research data available to others in a suitably confidentialised and well documented form.

Other benefits from a scientific and ethical perspective to the preservation and reuse of social science data for secondary purposes include:

- Provision of a secure environment for the management and storage of research data, and potentially also for the analysis of such data (see further discussion on data access environments below);
- Supporting appropriate levels of access to data for both researchers and participants - consistent with the secure access and digital repatriation capabilities identified in the Issues Paper for health and medical data (Sections 5.2.5 and 5.3.2); and
- Enabling the reuse of data to limit burdens on research participants and the community for the collection of new data, and also potentially enables the access to additional benefits from the research, through (appropriate) access to the research data as an output of the research process – consistent with the "Justice" guidelines of the NHMRC/AVC.[13]

[13] NHMRC/AVCC (2015) National Statement on Ethical Conduct in Human Research, 2007 (Updated May 2015). Available from: https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72_national_statement_may_2015_150514_a.pdf

Access to research data is also consistent with recent developments in the area of reproducible research in the social science community and open science programs more generally. For example:

- Center for Open Science in the USA has established their "Open Science Framework" with a mission to "increase openness, integrity, and reproducibility of scientific research"[14];
- Data Access and Research Transparency[15] (DART) initiative established in the US political science community focuses on increasing transparency in social science through the implementation of a set of policies and practices, including through journal publishers and editors; and
- American Economic Association now has a stated policy "to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication".[16]

These approaches have ***access to research data***, along with ***research publications*** and ***statistical analysis code***, as core elements of their transparency and reproducibility models – which has implications for the provision of data infrastructure facilities to support such models.

2.2 Public trust

Each of the above initiatives aims to improve social science research practice. Notably however, they have the additional benefit of enabling public trust through demonstration of the transparency of methods and practices used within different social science research communities. They also are encouraging a new generation of researchers committed to sharing research data, code and publications as a core principle of good social scientific practice.

The importance of public trust in social science cannot be understated. As the recent public concerns raised over the collection of the 2016 Australian Census have highlighted, trust in data collection and access, and the agencies that support them, is critical to enabling social science research and practice to occur. Consistent with this trust model, social science data archives have increasingly been accrediting through a trusted digital repository model, the Data Seal of Approval, as a means for enabling trust in the storage and preservation of research data among researchers, funders and the community at large.

14 https://cos.io/about_mission/

15 http://www.dartstatement.org

16 https://www.aeaweb.org/journals/policies/data-availability-policy

2.3 Return on public investment

Social surveys and qualitative research are widely used in Australia for academic research purposes, evaluation of government policies and programs, ascertaining community attitudes and to help inform the development of policy particularly in the areas of social policy, health and education. Many millions of dollars are spent conducting social surveys and undertaking qualitative research in Australia to support either basic research or to inform policy development and evaluation. It is also becoming increasingly difficult and costly to collect primary survey data.[17] However much of the survey data collected is underutilised.

The reuse of secondary data increases the return on the investment in the initial data collection, but can also mean that it is not necessary to undertake a new survey and thus reducing primary data collection costs. The costs of high quality surveys can be very high. For example, the cost of the four longitudinal studies currently being managed by the Department of Social Services[18] is approximately $300 million since 2001. These costs make it imperative that data producers to avoid duplication of research data collection efforts – both for efficiency reasons and to minimise the burden on the communities and populations being studied.

The other notable characteristic about the return on investment from secondary data use is that it is ongoing – the data does not degrade, and thus is able to be reused over and over again by an increasing number of researchers. It further becomes increasingly valuable when linked with other data sources – such as medical or education records to provide policy outcomes data, and with census data to provide contextual data. This reusability (and non-excludability in economic terms) makes the research data itself an infrastructure asset – distinct from the services using those data, which are often value-adding.

Finally, the availability of secondary data for social science research also expands the potential community of PhD and Masters students who are able to conduct research. Many social science research questions require access to data that is nationally representative, and often international in scope – and it is usually not possible to collect primary data to conduct such representative and comparative studies. Access to secondary data enables a significant volume of social science postgraduate research that would otherwise be impossible for cost reasons.

---

[17] There a range of reasons for this including falling response rates, the proliferation of marketing surveys and difficulties of engaging key groups such as disaffected young people in surveys.

[18] https://www.dss.gov.au/about-the-department/national-centre-for-longitudinal-data