Australian Research Data Commons

Humanities, Arts and Social Sciences Research Data Commons

Final Report

Alexis Tindall (ARDC), Ian Duncan (ARDC)

March 31, 2020

Australian Research Data Commons

Table of Contents

Executive summary

"The Humanities and Social Sciences are the study of human behaviour and interaction in social, cultural, environmental, economic and political contexts. The Humanities and Social Sciences have a historical and contemporary focus, from personal to global contexts, and consider challenges for the future.

Through studying Humanities and Social Sciences, [we] will develop the ability to question, think critically, solve problems, communicate effectively, make decisions and adapt to change. Thinking about and responding to issues requires an understanding of the key historical, geographical, political, economic and societal factors involved, and how these different factors interrelate.

The Humanities and Social Sciences ... provide a broad understanding of the world in which we live, and how people can participate as active and informed citizens with high-level skills needed for the 21st century."[1]

Humanities, Arts and Social Sciences (HASS) and Indigenous research is research by people about people. It underpins initiatives that aid community wellbeing and resilience in times of peace and upheaval, guides and informs government and community service decisions, drives innovation, creates new industries and effective regulatory frameworks and helps us to understand, benefit from, and preserve Australian and regional culture, history and heritage.

This report summarises the findings of consultations and analysis conducted between September 2019 and March 2020 across the HASS research, data and infrastructure communities to assess the HASS data landscape, research communities' priorities, activities and needs, and opportunities for leverage and enhancement in that environment. The report proposes a suite of activities to be rolled out over an initial five years to form the core of a HASS Research Data Commons, a key component of Australia's HASS National Research Infrastructure.

To be successful, a HASS National Research infrastructure investment must be broad-based, forward-thinking, innovative, and able to respond to evolving challenges. In arriving at a solution that enables innovative research today, overcoming contemporary limitations, with an eye to the future, the core question is: "What foundations, laid today, can be built on for success tomorrow?"

This report envisions infrastructure that enables skilled digital researchers, working in multidisciplinary teams, to use and contribute to the development of the best digital methods to generate new insights into society and culture while providing reliable and effective forecasting on key future social issues, strategies, and outcomes.

HASS researchers should have access to complete, thorough, timely, organised and informed data about today's society and digital access to valuable collections locally and internationally. Integration of data and computing

[1] https://www.rossmoyne.wa.edu.au/programs/learning-areas/humanities-and-social-sciences/

services will promote efficient and innovative research into our past and future, providing a sound and trusted base for informing policy and service delivery decisions, driving benefits for other domains and society.

A Research Data Commons (RDC) improves the breadth and efficiency of research, amplifies the scale of emerging research methods and approaches, encourages and supports responsible research and creates new interdisciplinary and transdisciplinary research and translation opportunities.

In essence, a Research Data Commons is successful when it provides researchers with a global competitive advantage through access to and use of data.

A HASS RDC (or one for any research community) involves the successful integration of cultural, social, and technical solutions with policy, governance, training and skills complementing and enabling world class research.

The benefits of investment in a HASS Research Data Commons include improving access to and management of data from and about Indigenous communities, empowering community decision making, and broadening ways for Australian researchers to work in partnership with our Pacific neighbours. This investment will provide the HASS community with new tools and techniques to improve the quality, timeliness, range and depth of advice provided to government in the pursuit of data-driven economic, social, and cultural policies, increase transparency and robustness in relation to publicly funded research and foster and encourage the research leadership that grows around accessible and useful datasets and cutting edge research infrastructures.

Current state

Consultation revealed that HASS research communities include a broad range of disciplines with great diversity of activity, approaches and readiness to benefit from investment in a research data commons. This community draws on diverse data sources, including:

- Born-digital and digitised government data (contemporary and historical),
- Gallery, Library, Archive and Museum (GLAM) collections,
- data generated through research, including survey and sensor data,
- data created by instruments,
- audio-visual material, and
- data generated and managed by commercial providers, including social media and other platforms used by the broader community.

Despite the diversity of approaches and take up, the impact of data-enabled research and digital research methods is evident across these communities. In some cases data-enabled research is driving new methodological approaches, including the emergence of sub-disciplines and transdisciplinary collaborations. In other cases researchers are using data and new tools to expand the capacity of, or to accelerate, traditional methods. HASS communities are benefiting from the emergence of large datasets and accessible tools, including visualisation and analysis tools.

Despite the rich data landscape and the adoption of new data, tools and methods, consultation demonstrated that the landscape is fragmented, and the benefit of these tools and data access is in many cases limited by the absence of national infrastructure or coordinated integration of existing resources. Additionally, the uneven distribution of this work limits collaboration and slows the development of new research communities.

Opportunities

This report identifies a range of activity in the local and international HASS research landscapes that can be leveraged or enhanced as part of development of the HASS Research Data Commons. This includes, but is not limited to, contemporary effort dedicated to sharing government data, and collaborative work with other National Collaborative Research Infrastructures, where common issues can be identified. It will work with a range of established and newly announced institutional and LIEF-supported initiatives that make up the current de facto national infrastructure, the digitisation initiatives of the gallery, library, archive and museum (GLAM) community and capitalise on emerging big data in HASS relevant collections.

Overview descriptions are provided of a range of relevant initiatives, and a collection of nationally significant collections are characterised as part of this report.

Key focus areas

The six focus areas that inform the HASS Research Data Commons proposal summarise key principles identified through consultation and align with Commonwealth Government initiatives:

- **Aggregation of and access to existing data**
- **Improved analysis environments for aggregated data**
- **Data linkage for improved outcomes**
- **Secure linkage and analysis environments**
- **Indigenous Data governance**
- **HASS sector communication and collaboration**

The proposal recommends an initial cluster of activities that will demonstrate measurable impact within 12 months, with outcomes enabling the planning and execution of future developments over the five year time frame. Initial activities focus on “ready, willing, and able” groups with expansion into additional disciplines and communities being informed by the initial program.

The RDC will require large scale review at the five year horizon, where we anticipate researchers working at a national scale contributing to tools development, and access to high-quality data driving increasing impact as HASS research becomes more data intensive and models of sharing becoming normalised.

Overview of activities

Theme 1: Data

Objective: Improved access, re-use and enhancement of more high-quality data for researchers: particularly Government data, data sourced from the Galleries, Libraries, Archive and Museums sector, and other significant research datasets.

Benefits:

- Make extant data as useful as possible, and available for enhancement and enrichment
- Enable efficient and effective researcher access to increasingly available government data
- Prove the value of researcher access to GLAM collections at scale
- Enable new international links for Australian GLAM collections and associated HASS research, including:
    - improved access and use of our Pacific and regional collections, linking with stakeholder communities

- building links to expose Australian collections in relevant international initiatives, ensuring recognition of our cultural and material heritage in global environments

Activities:
Activities described under this theme include leveraging existing initiatives that are improving access to government data and investment in research platforms for GLAM collections demonstrating advanced state of readiness to provide researcher specific services.

Theme 2: Platforms

Objective: Demonstrate uplift in research practices and outputs through investment in collections, tailored interfaces and analysis tools.

Benefits:

- Model and test options for improved access to research collections and salvage of collections at risk, data linkage and integration with improved analysis environments.
- Support development of shared workflows, and collaborative digital research approaches.
- Leverage and enhance existing social sciences research infrastructure for improved outcomes through increase in scale.

Activities:
Activities described under this theme include development of a model research data commons, working with an initiative that exemplifies challenges and opportunities common to many HASS communities, as demonstrated through consultation, measures that will aid integration of existing research infrastructure and exploration of a framework for environments focused on the analysis of sensitive data.

Theme 3: Data governance, sovereignty and linkage

Objective: Focus on challenges that are particular and characteristic of HASS research data, with an interest in ethics and responsible development of data rich environments.

Benefits:

- Inform and complement other elements of this proposal
- Develop, test and implement practical responses to key challenges in data-enabled HASS research

Activities:
Activities described under this theme will advance implementation of Indigenous data governance and contribute to efforts for increased data linkage.

Theme 4: Community communication & collaboration

Objective: Developing strong, well-connected and collaborative HASS research communities, with an emphasis on communities and groups which are engaged in national or potentially national resource development and/or use.

Benefits:

- identification of potential shared services and resources such as:
- sharing and developing best-practices including

- initiating new collaborations

Activities:

Activities under this theme will support governance, collaboration and communication around the suite of activities.

Conclusion

The value of clear, fact-driven, and evidence-based strategies to address major challenges to society, the economy, and our place in the world has never been clearer than right now.

We have a generational opportunity to support and strengthen emergent research practices, building collaborative opportunities between Australian researchers and their international counterparts, maximising the value of public investment in research, and showcasing Australian research and collections for the world. Providing this investment at a national scale removes inefficiencies and empowers the HASS community to meet its potential in an increasingly digital global environment.

The success of a HASS Research infrastructure investment relies on it being broad-based, forward-thinking, innovative, institutionally supported, and able to respond to changing challenges. The task is to arrive at solution(s) that enable innovative research today with an eye to the future. We must think about tomorrow's researchers and what they will need to take advantage of and develop the research approaches of the future, based on and guided by the experiences, opportunities and issues we identify today.

Background

"[A Humanities, Arts, and Social Sciences (HASS)] national research infrastructure focuses on enabling inquiry across the research spectrum including research into cultures, communities, environments, health and social well-being. Humanities, Arts and Social Sciences (HASS) platforms range from physical collections across the humanities, arts, environmental and medical sciences to online portals that facilitate the digitisation of and digital access to original artefacts, materials and knowledge. In addition, HASS based platforms can be used to manage and integrate data to enable the development of solutions for complex social problems for the benefit of all Australians"

(2016 National Research Infrastructure Roadmap)

The 2016 National Research Infrastructure Roadmap (the Roadmap) identified opportunities that exist to accelerate the impact of HASS and Indigenous research through the improved overall coordination of research infrastructure supporting access to and analysis of physical and digital collections using tools such as digitisation, aggregation and interpretation platforms.

Specific potential benefits the Roadmap identified included opportunities to:

- Leverage existing portals and facilities
- Integrate HASS platforms with digitisation and next generation technologies
- Enable the improved multidisciplinary approaches that increasingly underpin the HASS sector
- Bring institutional capabilities collectively up to the level of national scale research infrastructure, while leveraging existing investment at an institutional level
- Enhance access to national and state collections and aid a greater degree of interoperability across collecting institutions
- Improve accessibility to physical items and build on digitisation efforts that are shaping the nature of HASS research
- Enable international interoperability with related initiatives and enable Australia to help shape international research infrastructure.

The Government response to these priorities highlighted an imperative to "explore better integration of information in the Humanities, Arts and Social Sciences (HASS) and Indigenous research data platforms to maximise research outcomes, across a large and diverse group of stakeholders" and committed funds over two years, commencing 2020-21, for a scoping study through Research Infrastructure Investment Plan (RIIP).

Investigation into a potential "HASS Research Data Commons (RDC)" was highlighted by a range of stakeholders in Department of Education, Skills and Employment (DESE) led consultations, who commenced the scoping of a HASS RDC as a specific potential national research infrastructure (NRI) capability as part of a wider HASS scoping study. The ARDC was requested to undertake this role as a leader in the space of NRI eResearch, specifically around the model of a data commons, and its historical experiences in the HASS Data Enhanced Virtual Lab (DeVL) project.

The purpose of this project is for the ARDC to provide advice to DESE on infrastructure gaps, needs and potential implementation costs of a HASS eResearch Data Commons infrastructure and associated activities to support improved capability in the HASS community by establishing a HASS RDC, with a particular focus on discoverability, accessibility, and interoperability of HASS data.

The ARDC has been asked to focus on the following key activities in scoping a potential RDC:

- Identifying potential demarcation lines to logically group communities of similar priorities and states of readiness
- Scoping the current state of interoperability for HASS data sets across those groups, including work already underway with Government and other stakeholders, e.g. data linkage and open data agendas
- Identifying the proposed foundations and goal of a HASS Research Data Commons (RDC).
- Identifying current gaps in discoverability, accessibility, and interoperability of one or more 'data commons' for driving HASS research in Australia, including governance, access, common metadata and usability of data. This should include what is and should be a national responsibility rather than institutional.
- Identifying a potential definition of what HASS data should be considered nationally significant for a purposes of a HASS RDC; consideration should be given to the work of the New Zealand Government.
- Mapping nationally significant HASS data collections across Australia and the location of their storage and management.
- Identifying opportunities to enhance and leverage existing NCRIS capabilities (i.e. ARDC, PHRN and AURIN) and other national and institutional facilities / projects, which are not currently funded by NCRIS, to underpin a HASS RDC.
- Providing advice and costs on implementing a consolidated HASS RDC, including staging, initial and ongoing costs and if the HASS RDC is one or multiple commons.

During the project the ARDC has engaged in consultation with many and varied stakeholders in HASS communities, including learned academies, representatives of research communities, higher education institutions, Galleries, Libraries, Archives and Museums (GLAM institutions) and other data providers to provide this advice.

Introduction

HASS and Indigenous research underpins initiatives that aid community wellbeing and resilience, informs government and community services, drives innovation, contributes to the creation of new industries and effective industry regulation frameworks, and helps us understand and preserve Australian and regional culture, history and heritage.

As Australia addresses the challenges arising during 2020, opening with our national bushfire crisis and continuing through the response to the global COVID 19 pandemic, accessibility of data and new models of research become more important than ever. Disciplines under the HASS umbrella include economics, education, law, public policy, behavioural science, communications and the arts, among others. Humanities and social science researchers contribute to how Australians and their leaders respond to crises. Real time response to crisis is aided by communications analysis, behavioural science and social media analysis. Planning, review and analysis can include new approaches to labour practices, considering the economic impact of proposals and decisions, new models of teaching and learning, exploring economic and geospatial perspectives on access to health services, and considering the cultural response to sudden changes to our way of life.

Platforms for HASS were highlighted as one of nine focus areas of proposed activity in the 2016 National Collaborative Research Infrastructure Roadmap, a recommendation that was accepted in the Australian Government's response with the comment that targeted focus will contribute to "strengthening Australia's economy, advancing societal benefit, improving our competitiveness and building on existing national capability" (Australian Government, 2017, p.11). In that response, the Government highlighted that national research infrastructure (NRI) "facilities are also melting pots for collaboration - they bring together researchers from across broad disciplines, catalysing relationships and supporting cross-disciplinary research. High quality and accessible NRI is a prerequisite for attracting international innovators and companies, driving Australia's reputation for scientific and research excellence. It attracts, develops and exports a highly skilled workforce, including the next generation of researchers" (p. 3)

Twenty years of project-based HASS investment has led to significant areas of strength in data-enabled research. Large, multi-contributor, multi-user collections have developed, some operating as de-facto national research infrastructures, and researchers and data providers are mining this landscape to power new forms of research that impact our understanding of ourselves and our place in the world. The expanding pool of current and potential sources of data adding value to research include Government data, GLAM collections, grey literature collections (material published by organisations whose primary purpose is not publishing but who are nonetheless a rich source of policy-relevant information), and social media.

This HASS research and data landscape represents considerable government and institutional investment and delivers world-leading outcomes that directly benefit Australia and Australians, as well as our regional neighbours. Building connections in that landscape, dramatically increasing the scale of data available, and improving the methods for using that data will supercharge the ability of HASS communities to continue to drive high-impact research that benefits all Australians.

The opportunity presented through the development of a HASS NRI is therefore to leverage these foundations and develop the integrations, coordination and coherence that will push the frontiers in transdisciplinary research, ensuring that the next generation of HASS researchers have the skills, tools, and environments needed to develop and support a strong, agile, and innovative Australia.

What is a Research Data Commons?

"Data commons collocate data, storage, and computing infrastructure with core services and commonly used tools and applications for managing, analyzing, and sharing data to create an interoperable resource for the research community"[2]

Why a HASS Research Data Commons?

As the availability of data increases, and methods and practices mature, the needs of HASS research communities extend beyond the capacity of individual researchers and their desktop computers and it is proposed that a Research Data Commons (RDC), based on the FAIR principles (see Appendix 1), will improve the efficiency of research, amplifying the scale of emerging research methods and approaches, encouraging responsible research, and creating new interdisciplinary and transdisciplinary research and translation opportunities.

A Research Data Commons (RDC) for HASS (or any research community) involves cultural and social solutions as well as technical solutions. Policy, governance, training and skills complement and enable world class research environments and a mix of technical and social activity is required to enable this research transformation. The mix includes development of and access to:

- underlying infrastructure;
- platforms and tools;
- frameworks and standards;
- policy and incentives;
- community and culture development; and
- measures to develop a skilled workforce.

Such a Research Data Commons, built from reusable tools and techniques, has the potential to form a critical component of the national HASS research infrastructure and will build on existing tools, standards and frameworks for data sharing, adding value to past investments and promoting and enabling innovative new research activities. Several research communities in Australia and internationally have demonstrated the potential of the data commons model to enhance and transform research practice and outcomes. HASS can therefore readily learn from and re-use the experiences of the astronomy, geosciences, marine and biosciences research communities in the development and use of their commons environments, while building new capability where challenges emerge that are unique to the HASS community.

In addition, international models demonstrate that new research opportunities become possible through the availability of data at scale, including explorations of intergenerational social and income mobility through generations of population register data[3], mapping of new measures of inequality[4]. High quality and accessible research infrastructure for the humanities, arts and social sciences will attract, develop and export a highly skilled research workforce and act as an incubator for the next generation of research leaders. Access to large, well defined and well organised datasets and alignment with new tools will result in further research skill

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636009/

[3] Explored in Denmark in Landersø, R and Heckman, J.J. 2017 "The Scandinavian Fantasy: The Sources of Intergenerational Mobility in Denmark and the US", in *Scandinavian Journal of Economics Jan; 119(1): 178–230. doi: 10.1111/sjoe.12219*

[4] Demonstrated in the MIT Lab Atlas of Inequality.

development and future innovative research. Access to more and higher quality data will enable more flexible and rigorous pattern recognition, improved correlation and connections across datasets, better visualisation using data, and statistical analysis using these datasets.

Benefits of investment in a Research Data Commons (RDC) tailored to HASS communities include:

- contributing to step-changes in HASS research for multiple disciplines through:
    - increased availability of high-quality data, particularly for computational analysis;
    - development of new platforms for collaboration and sharing
    - enabling the development of innovative methodologies
    - recognition of digital outputs and digital research excellence as first class research outputs;
- preservation and aggregation of data that might otherwise have been lost
- maximising the value of existing investment in HASS data and research across multiple sectors through:
    - increased accessibility and utility of digitised and born-digital outputs from the GLAM sector;
    - more efficient and effective access to government data; and
    - opportunities for the emergence of new research communities and transdisciplinary methods
- testing and applying approaches to the implementation of appropriate governance of indigenous data, making data about and collected by Indigenous communities as open as possible and as closed as necessary
- connecting Australian researchers and collections with international infrastructures, creating new collaboration opportunities and encouraging global awareness of Australian research and activity
- improving opportunities for Australian researchers working on challenges that impact our region, including the ability for Australian collections and research to work in partnership with our Pacific neighbours
- providing the HASS community with new tools and techniques to improve the quality, timeliness, range and depth of advice provided to government in the pursuit of data-driven economic, social, and cultural policies
- increasing transparency and robustness in relation to publicly funded research and providing frameworks to share data responsibly, contributing to a changing culture around shared data
- fostering and encouraging the research leadership that grows around accessible and useful datasets and cutting edge research infrastructures.

A HASS RDC investment will allow the Australian HASS research community to continue to contribute to cutting edge international HASS developments and foster and encourage the research leadership that grows around those datasets and infrastructures. Innovative researchers should be applying their skills to Australian challenges, and investment in coherent and integrated infrastructures will, in many cases, remove research limitations and allow researchers to work with the best data from the best sources.

Development, integration, and coordination in pursuit of a fully formed HASS Research Data Commons will be the culmination of a multi-year, multi-stage and multi-discipline investment and in this report the ARDC provides a summary of the current landscape, themes and priorities emerging from many consultations, a proposed infrastructure state, and proposals regarding a suite of activities that form a pathway towards that state.

Investment in effective underpinning research infrastructure is imperative as we look to the future of HASS research. Opportunities to develop new methods and approaches to key humanities, arts and social sciences questions become available as the challenges created by fragmentation and dispersal, limited access to large datasets, and outdated tools and infrastructures are addressed. As we examine and learn from large scale HASS infrastructures overseas we also recognise that such environments do not happen immediately; they are built on foundations developed by researchers and their support communities in strategic and multi-year programs.

HASS Research Landscape - current state

HASS Research Communities

For the purposes of this project, Humanities, Arts, Social Sciences and Indigenous research (HASS) are defined as the disciplines represented by Divisions 12 through 22, inclusive, of the *Australian and New Zealand Standard Research Classification (ANZSRC), 2008, Field of Research Codes.*

Reflecting one of the challenges facing a HASS RDC, this definition represents a very broad range of activities, clustering research as diverse as Urban Environments and Design, with Law, Economics, Classics and the Philosophy of Religion.

Furthermore, HASS research communities draw from a range of data sources, administered in a range of environments within and outside institutions. These include:

- Born-digital and digitised government data (contemporary and historical),
- Gallery, Library, Archive and Museum (GLAM) collections,
- data generated through research, including survey and sensor data,
- data created by instruments,
- audio-visual material, and
- data generated and managed by commercial providers, including social media and other platforms used by the broader community.

In some cases, HASS disciplines use the same large collections of data in different ways, while in other instances a data collection or aggregation of collections may only be of relevance to a small number of disciplines.

In some cases, data-enabled research is driving the development of sub-disciplines and/or new methodological approaches. This can be seen with the emergence of Learning Sciences as a transdisciplinary approach in schools of education and the use of gaming in philosophy and sociology to experiment with, and teach logic and reason.

In other cases, researchers are using data new tools to expand traditional methods, as demonstrated by historians' use of the Trove cultural collection, or to experiment with new methods, as seen in visualisations of the Tasmanian Convict Archive. The availability of larger datasets and novel tools for analysis is enabling new discoveries in life course studies - research that can demonstrate impact of life experiences on the health, prosperity and life expectancy of a person and their descendants - an area that can have significant impact on policy.

Support for data linkage, or the connections between datasets, is also making new insight into health, education and economic policy possible, for example by using geospatial modelling to explore the spatial distribution of determinants of variations in health outcomes between communities, including cultural factors, mobility and health provider access[5].

It therefore becomes apparent that determining priorities and states of readiness of potential participants in a HASS RDC *on the basis of discipline* quickly becomes complex and of limited usefulness, especially when it is

[5] Described in AURIN Case Study 'Hepatitis B Treatment Access'

considered that in some cases HASS researchers conducting data-enabled or data-driven research are atypical of their field. Some methods are common across communities, some researchers are isolated in their discipline, and some of the data-driven research demonstrating the greatest impact is multidisciplinary or transdisciplinary. For example, methods such as visualisation, named entity recognition, geocoding and mapping are common across many disciplines.

We have identified several potential approaches to grouping communities experiencing similar priorities, regardless of discipline, including on the basis of:

- **Methods and tools**: Common tools are evident across disciplines, including analysis tools such as Jupyter Notebooks and R, and are used in HASS and other domains for the analysis of "big data". Similarly, network analysis and visualisation tools are used in diverse applications across research areas. These tools are often not unique to HASS disciplines but theoretical approaches and training needs are HASS-specific. Such methods and tools could be supported by common approaches to skills development, provision of computing resource, collaboration and analysis environments and publication platforms.
- **Type of data**: HASS-relevant data is extraordinarily diverse and there are several ways of 'slicing' the community's current and potential data assets including but not limited to:
    - **Qualitative or Quantitative data**: Quantitative data can be counted, measured, and expressed using numbers while qualitative data is descriptive and conceptual. Qualitative data can be categorised based on traits and characteristics. Quantitative data analysis is common in social sciences, and emerging in some other humanities fields with increased access to large datasets and this distinction is a key differentiator across HASS data-enabled research communities and has considerable impact on their approaches to data sharing and re-use, and the skills required to analyse that data. For example, qualitative data underpins research from communities such as Education, Indigenous Studies, Creative Industries, and Anthropology. Some researchers in these fields take a risk averse approach to sharing qualitative data while others have come together to develop practical approaches to sharing and collaboration, as demonstrated by the ARDC supported Studies of Childhood, Education and Youth project.
    - **Text, images, & complex data**: A process of clustering researchers based on consistent types of data regardless of discipline can guide effective allocation of resources with respect to computing and analysis resources and environments, and skills development and advocacy, including policy development around ethics. Researchers can be supported based on the skills, environments and platforms that they need to make the best use of their data which could be access to digitised collections using the International Image Interoperability Framework, improving integrity of image sources, and limiting reproduction fees or other limitations on research use. It could equally be support for integration of data-level access to text heavy collections with accessible online text analysis tools, thereby expanding options for researchers without complex coding skills. Or it could be brokerage to support development of interoperability frameworks for emerging types of data, such as augmented reality or laser scanning data.
- **Data sources**: HASS relevant data are often sourced from Gallery, Library, Archive or Museum collections (GLAM) and government data, including surveys and administrative data, or are the product of research, including interviews, surveys, sensors, observations and the like. A HASS RDC will improve accessibility and useability of extant data such as digitised collections or government-sourced data; will develop options for publication and preservation of new data, will promote considered approaches to ethics and embargo periods; and will facilitate or create environments to manage complex data, including 3D, laser scanning, VR and AR material.

A Research Data Commons approach allows us to leverage existing Data Commons experience to plan a research infrastructure investment that accommodates a range of activities, a diversity of approaches to data and sources, and differing research ambitions. It can allow the development of individualised interfaces and

environments that meet the needs or preferences of disciplines grouped in multiple ways, built on an interoperable and sustainable substrate, ensuring that connections with major data providers in the landscape are reusable and adaptable.

Experience has also shown that initial RDC investment is more likely to demonstrate rapid benefit for communities with a higher degree of readiness, with subsequent expansion and translation of RDC capabilities into other disciplines being informed by the success or otherwise of these initial activities.

Other Stakeholders

HASS research and data collections also involve, in many cases, stakeholders outside the research community. Instances that have been outlined in the consultations include:

- **Subjects and beneficiaries of research*:***
    - **Minority language communities**; who contribute to, and use recordings of languages other than English in Australia. Similarly, considerable Australian holdings of Pacific languages are used by communities and researchers across that region.
    - **Indigenous communities and individuals**; discussed in greater detail in the next section of this report
    - **Interviewees or other subjects of research**; especially qualitative research.
- **Government**; considerable government research is supported by these collections, with the Australian Data Archive, AustLII, AURIN and Analysis and Policy Observatory reporting a high proportion of government users.
- **Public**; some HASS data collections are committed to ensuring public access, including AustLII, which describes free, anonymous access to law as a community justice issue, and AustLit, which reports state, school and community library subscription to their collection of Australian literature.

Consideration of these stakeholder communities will impact the development of a Research Data Commons in several ways:

- New forms of access to aggregated collections, particularly beneficial for minority language communities and other communities that may reconnect with historical research about their culture or antecedents.
- Maintaining and expanding public access for citizen users, even considering avenues for encouraging and capturing their contributions. Examples of this can be seen in transcription of handwritten or otherwise inaccessible sources[6], and new opportunities to responsibly capture and store Indigenous traditional knowledge[7].
- Strategic consideration of how to apply the FAIR principles to collections with complex access requirements, including:
- Qualitative data, which may include personal information about research subjects and can pose particular challenges with regard to de-identification
- data that is subject to community ownership, particularly data collected from and in partnership with Indigenous communities

[6] See the UTS Centre for Australian Public History's *Criminal Characters* project

[7] Including Mukurtu, in use by the State Library of NSW and Jumbunna Institute for Indigenous Education and Research, and Ara Irititja, a mature project developed by the Ngaanyatjarra, Pitjantjatjara and Yankunytjatjara people in partnership with South Australian collections organisations.

- Recognition that new forms of aggregation, access and linkage may risk making non-sensitive data collections sensitive, particularly if linkage exposes personal or sensitive information.

Solutions to aggregate or improve access to collections for research must complement and/or sit alongside concepts of accessibility of collections for other users. Similarly, designers and those involved in RDC governance must maintain a commitment to the best interests of all stakeholder communities. This is in line with Principle 5 of the Australian Code of Responsible Research "Respect for research participants, the wider community, animals and the environment", which is expanded upon with the comment "treat human participants and communities that are affected by the research with care and respect, giving appropriate consideration to the needs of minority groups or vulnerable people".

Indigenous data

Any initiative to aggregate and create access to HASS relevant data on a new scale or through new methods will inevitably need to take a responsible approach to Indigenous data.

Over recent years, researchers in Australia and other countries with similar histories have considered models to address the challenges of data sovereignty that have arisen as a result of colonisation. Approaches that have emerged include the CARE Principles of Indigenous Data Governance, a model that aims to ensure collective benefit from research and appropriate approaches to authority over data and ethics; and Indigenous Data Sovereignty, which affirms the right of Indigenous peoples to govern the collection, ownership and use of data about Indigenous communities, peoples, lands, and resources. This sovereignty is implemented through Indigenous data governance.

Ongoing efforts to respond to these challenges, including the work of AIATSIS, their Guidelines for Ethical Research in Australian Indigenous Studies (presently under review), and other commitments by relevant organisations, will benefit from complementary work to develop community-agreed models to collaboratively manage an Indigenous Data Governance framework in Australia.

HASS research is research about people, which draws upon and creates data about people. Sensitivity around Indigenous data will not necessarily be confined to that which is flagged as Indigenous research. GLAM organisations are involved in efforts to 'decolonise' their archive, re-examining methods of collecting and describing Indigenous cultural material and traditional knowledge[8]. Libraries and archives are exploring methods of improving cultural safety for Indigenous people and others using their facilities[9][10], working on improvements to archival practice in line with the cultural change that has taken place in healthcare provision in recent years.

There is therefore an opportunity for Australian research to benefit significantly from ensuring indigenous knowledge management and indigenous knowledge governance are core components of the aggregation of datasets from diverse sources, maximising its potential to build on existing Eurocentric knowledge management practices, and reinforcing individual providers' attempts to improve cultural safety. Incorporation of a broader view of knowledge management and governance will also avoid the creation of practices which limit research capability, for example risk-aversion around new linkage opportunities making non-sensitive data sensitive.

[8] Australian Museums and Galleries Association, First Peoples: A Roadmap for Enhancing Indigenous Engagement in Museums and Galleries
[9] Monash University Statement of Principles relating to Australian Indigenous Knowledge and the Archives
[10] National and State Libraries of Australasia, Culturally Safe Libraries

All HASS data, and many forms of HASS research, may include consideration of Indigenous people, or data sourced from Aboriginal and Torres Strait Islander communities, so the design and implementation of a HASS Research Data Commons should support the principles that:

- Activities reinforce and support the investigation and implementation of community-agreed Indigenous data governance practices.
- Activities are informed by a strategic and sensitive approach to managing and sharing HASS-relevant data about and owned by Indigenous people, referring to international resources and developments, including the work of the Indigenous Data network, relevant networks of the Research Data Alliance, and the CARE Principles of Indigenous Data Governance.

The current HASS Data Landscape

As we have outlined, HASS researchers work in a rich data landscape, using data drawn from diverse sources, as well as generating and sharing their own research data. Through individual and small group consultations we have collected information about numerous data collections in use by, or of interest to, the HASS research community (listed in Appendix 3). These consultations have explored the nature of the collections, their location and storage, discoverability, accessibility and interoperability. Data providers have also shared observations on user communities, potential for the future of their collections, and we have discussed current and future challenges including their sustainability.

It is apparent that essential components of a successful RDC are the discoverability, accessibility and interoperability of data and analysis assets and in these consultations we have explored the nature of the collections using the FAIR framework.

Our consultations have illustrated a rich and diverse data landscape with great potential for additional value and impact through addressing the challenges of fragmentation and accessibility that impact the research potential of these collections.

Accessibility is a core FAIR principle and it is evident that **highly accessible** collections drive research and collaboration, lend themselves to linkage and aggregation, and make international connections and aggregation possible. While HASS, in common with many research domains, displays a limited number of collections that demonstrate the highest implementation of the FAIR data principles, accessible is used here to indicate some degree of findability, accessibility, interoperability and reusability. Conversely, **inaccessible** collections result in limited or missed research opportunities and, through lost research or the ongoing maintenance of collections that are not widely available for reuse, are a poor use of public funds.

The Australian HASS landscape includes collections of national significance that sit in between these points, and could be activated for greater impact by inclusion in a Research Data Commons model.

Cutting edge data-driven or data-enabled research depends on availability of data at scale. Data-driven research often depends on access to ‘big data’, while some will draw on small, valuable and niche data. In many cases the instances of big data in use in HASS research are aggregations of small data. Even small datasets benefit from availability of data at scale, which can influence discoverability, provide opportunities for linkage, promote a culture of sharing and enable new forms of research.

Observations, clustered according to “accessibility”, are summarised below. Detailed information on individual data providers and collections are shared in Appendix 3.

Data Categories:

Highly accessible HASS data collections:

Definition: *Data and metadata are discoverable, accessible through standard communication protocols, community accepted metadata and vocabularies are used, content is licensed to indicate suitability for re-use, high quality application of the FAIR principles.*

There are few instances of Australian HASS-relevant data achieving the highest implementation of FAIR data principles. There is appreciation of the value of using community accepted metadata and vocabularies and reasonable implementation of that to aid discoverability, but elements that facilitate machine readability to ensure computational access and utility of collections are in limited use. The benefits of computational accessibility lie not just in researchers’ ability to efficiently find and retrieve data and analyse it using digital tools, but also in ease of aggregation, building international links, and integration with platforms and other analysis environments.

Examples in the Australian HASS research environment:

- Trove: Content from Trove’s collections (including digitised newspapers and gazettes, journals, articles and datasets, books, theses maps, diaries, letters and archives etc) is accessible through an Application Programming Interface (API), allowing machine to machine communication with data about people and organisations being accessible through a separate API. These APIs help users use Trove data in other analysis environments or offline, create new tools and visualisations, and draw Trove data in for display in other web environments, among other things. These APIs have enabled:
    - To Be Continued, research project used paratextual search of digitised text from Australian Newspapers to identify more than 9000 previously unknown items of extended (greater than 10,000 words) nineteenth century literature.
    - Data from various Trove collections to be drawn into accessible computational analysis tools such as Jupyter Notebooks and R. Examples of Jupyter Notebooks using this tool were developed and shared in the Tinker Studio, as a model for humanities researchers new to these methods.
    - Harvesting of Australian data into Europeana 1914-1918, providing a federated search across European, American, New Zealand, and Australian collections relating to the First World War.
- Australian Urban Research Infrastructure Network (AURIN): AURIN, an NCRIS initiative, is an online workbench with access to thousands of multi-disciplinary datasets, from hundreds of data sources and analytical tools covering spatial and statistical modelling, planning and visualisation. Providing researchers with access to diverse sources of data, the ability to integrate data across disciplines and interrogate that data to answer their research questions, AURIN is an invaluable resource for researchers and government exploring demographics and social indicators, urban design and planning, housing, health and livability, infrastructure, transport and economics, among many others. Integrations with training resources, mapping and visualisations tools, and an excellent API means that computational access to data is well supported. The API is built on standards from the Open Geospatial Consortium, and facilitates integration of AURIN resources with analysis tools (including GIS tools, python, R and Jupyter), mobile apps and web services.
- Australian Data Archive: Built on the internationally-used Dataverse platform, uses standards in line with international best practice, and is actively involved in international interoperability initiatives to enable sharing social sciences data across platforms. Provides API access, including access to sensitive data. This has made it possible to share historic census data from Australian Data Archive to the geospatial environment of

AURIN, where it can be combined with a range of other relevant geospatial datasets. Newly available census data extended the range of longitudinal analysis using the AURIN Portal.

- PARADISEC: Enables robust description of multimedia collections and is interoperable with tools used in linguistics environments, including Elan annotation services and and Fieldworks Language Explorer for documentation and analysis.
- PARADISEC Metadata is structured and exposed so that it can be harvested by Open Languages Archive Community, a massively successful example of linked open data providing a worldwide virtual library of language resources.

Collections with limitations on use

Definition: *Discoverable and well organised, widely used sources of data that have restricted access models (eg technical limitations, subscriptions, licensing, etc)*

Despite the high degree of functionality of these data collections, and their high level of impact on research practices in relevant disciplines, their potential for transformative research and their ability to power innovation is limited. Limitations can include lack of development resources to make data accessible in ways that can drive new research, licensing limitations, or subscription models implemented to aid sustainability that then limit openness. Involvement in a Research Data Commons can broaden the potential of these significant collections to drive new research.

Examples in the Australian HASS research environment:

Australasian Legal Information Institute (AustLII): Indispensable resource for legal and legal research communities, as well as the general public, AustLII provides free, anonymous, digital access to law, case law, legislation, legal commentary, law-related publications and other material including reports around Royal Commissions and related material. Data is organised according to community accepted standards and AustLII works closely with Legal Information Institute (LII) services from other countries, to the extent of the Australian instance hosting equivalent services for New Zealand and Pacific Countries. Structure of the facility is in multiple, related platforms which limits search functions and the ability to provide integrated access across layers. Much of the hosted material is subject to Crown Copyright, meaning that full open access is not possible, and users seeking reproduction of the material need to deal directly with the Courts in question. AustLII is, however, interested in supporting researchers with improved analysis environments (suitable for sensitive data) which are beyond the scope of their current facility. They are interested in building machine learning capacity, integrating natural language processing tools to enable automatic summary, and enabling analysis using symbolic logic.

AusStage: Transformative access to history of performance in Australia, linked data of people, works, theatres, organisations drawn from ephemera held in physical collections. Built on Dublin Core metadata schema, data model is described, all identifiers persistent. Allows programmatic access using a variety of web services that enable harvesting by Trove, AustLit and HuNI. Basic network analysis tools are provided as part of the AusStage interface, but this is an area of extraordinary potential for that data if they had development resources to adapt their infrastructure to provide full access for network analysis.

AustLit: A searchable, scholarly source of authoritative biographical, bibliographic, critical, and production information about Australian writers and writing, including more than one million records, and 1000 full text works, as well as links to where many other full texts are available online. AustLit's BlackWords provides access to an unparalleled record of Aboriginal and Torres Strait Islander Writing and Storytelling. In an effort to aid sustainability, AustLit has implemented a subscription model for accessing their data. AustLit is functionally accessible to researchers, as many university, state and territory libraries maintain subscriptions, but this model

prevents some forms of computational analysis as they are unable to make their data open without threatening their financial viability.

Population Health Research Network (PHRN): NCRIS supported PHRN is a national collaboration linking life data, generally data collected through the provision of health and human services, by federal, state and territory governments, private and not-for-profit organisations, to support research focussed on health and wellbeing. Supported by state and territory governments and academic institutions, the PHRN draws together the work of dispersed data linkage units, and makes linked data available either by secure file transfer or in secure environments, including the Secure Unified Research Environment (SURE), a remote-access data laboratory operated by the SAX Institute. In this instance there are considerable and reasonable restrictions on access, for important privacy and ethical purposes. Researchers seeking to access linked data for cross-jurisdictional and multi-jurisdictional research projects are required to apply for approval, which is considered by data custodians, linkage units and a Human Research Ethics Committee. Important research is supported by PHRN services, which is a trusted and proven model of access for sensitive linked health and human services data. Researcher consultation for this project has demonstrated broader interest in linked data than that which is provided by PHRN, and dissatisfaction with the cost of accessing SURE. PHRN and facilitators of the SURE environment can provide useful expertise in the extension of linked data services, either through extending exsting services or informing the development of new services.

Collections with limitations on Discoverability

Definition: *Collections of data that may be well described, highly organised and accessible but are not made discoverable in such a way as to make the data researchers seek easily found.*

National Archives of Australia: The National Archives (NAA), a primary source of information for the significant historical Australian Government data, can be described in all three categories of highly accessible, limited accessibility and inaccessible data, all at the same time for the same data! A mature digitisation program means that many records are fully available through their RecordSearch function and the organisation has produced many discoverability aids that are very accessible. The NAA also provides a digitisation on demand service for a fee based on cost recovery. Accessibility is high for those records that are digitised, an example being that historian Tim Sherratt was able to build tools to harvest data that are shared through his GLAM workbench[11]. However, the metadata and schemas used to describe the data is often shaped around an administrative agenda which creates complications for researchers. An example could be records described by their administrative function ("Department of Immigration and Multicultural Affairs, series 1 through 36, 1995 to 2005"), when what the researcher wants is 'Management, administration and closure of the Woomera Immigration Detention Centre, ministerial correspondence, submissions received leading to its closure". Historically such an issue has been resolved through the production of finding aids. The National Archives delivers many finding aids on their website and in published guides to records, but an effective integration of this approach with the digital environment has not yet emerged. During consultations researchers have described the resultant inaccessibility of files with metadata which is not “researcher-ready”, as well as the impact of the costs of digitisation on the nature of their research. In 2020 the National Archives is involved in a large scale review of information management, which may benefit from integration with a new Commons activity.

Institutionally held collections: Many significant research data collections are managed in institutional repositories, with varying degrees of access. An example of this is the Mitchell and Delbridge recordings for study of Australian speech. This database contains recordings of Australian English as spoken by 7736 students at 330 schools across Australia, mostly collected in 1960. This is highlighted in this report as an example of an institutionally held collection that would benefit from involvement in a Research Data Commons as this

[11] Described at GLAM Workbench/ RecordSearch

collection is of high interest to national initiatives such as the Languages Data Commons of Australia (LaDaCA, described later in this report) but is presently only available through the University of Sydney Library, limiting its potential for aggregation and new forms of analysis. Through the earlier work of the NCRIS ANDS project, many institutionally held collections are now discoverable through Research Data Australia but analysis of the accessibility of those collections (conducted as part of the Tinker/ HASS Data Enhanced Virtual Laboratory project) indicated that none of the 35,000 collections flagged as relevant to HASS were machine actionable (i.e. open, appropriately licensed, and providing a direct link to data).

Inaccessible collections

Definition: *Refers to a range of challenges including collections that are not digital, data that may not be lodged in a repository, are subject to uncertainty over ownership and reuse requirements, or where risk-averse approaches to sharing data have limited their reusability. This is not an exhaustive list of challenges in the HASS data landscape, but represents matters that have been raised in consultation to date.*

Physical collections that remain digitally inaccessible

Limited digital access to GLAM collections is an inhibitor to research. The extraordinary value of Trove itself, as a nation-level collection, is well articulated and accentuates the potential benefits from increasing the proportion of holdings of Australian galleries, libraries, archives and museums on both national and state levels that are digitally accessible.

Where metadata about physical GLAM collections are digitally discoverable, researchers have reported limited catalogue metadata preventing discovery of data, and the challenges of accessing data including full text transcription, audio-visual files or other multimedia. They have also described impediments including costs to digitise materials, or in one case, to access collection lists on site, that have either been prohibitive or have had an impact on the scope of their research.

Researchers are acutely aware that Australian GLAM institutions are progressing with the digitisation of their collections within both their resource availability and mandates, given it is not explicitly the function of GLAM organisations to support research, and researchers are interested in ensuring that the value of that digitisation effort is maximised. They are therefore keen to have input into digitisation priorities and the value of digitised material would be amplified if we could improve accessibility through aggregation or platform services which, in turn, would lead to improved decision-making around future digitisation priorities.

AustLII has also expressed interest in further digitisation, as many law resources still only exist in paper format, and data around the law has very long term value.

Lost data

Several researchers have expressed concern about the loss of potentially significant data collected during ARC-funded projects prior to widespread institutional support for data management, and which now lives only on researcher hard drives, floppy disks, or has been lost. There is potential to rescue “lost” data and current efforts by national funders to compel improved data management practices and institutional repositories aim to improve the ongoing value of data and limit the risk of future loss. Establishment of an RDC that contributes to the normalisation of high-quality data sharing practices, can measure the impact of reuse, and integrates with persistent identifier systems, ultimately increases researcher career recognition for “good” data sharing and may deliver the cultural shift to further reduce the risk of data loss.

Examples of lost or inaccessible data have been shared by the linguistics community, including where an approach to a retired researcher leading to the re-discovery of many 40+ year old recordings that linked to a contemporary project, greatly enriching the research outputs. PARADISEC are also able to share stories of the retrieval of researcher recordings of at-risk languages, sometimes on near-obsolete media.

Risk aversion restricts sharing of qualitative data

A number of HASS disciplines work with qualitative data (non-numerical or unstructured data collected through interviews, observations, focus groups and the like). Researchers consulted in Creative Industries, Education and Indigenous Studies expressed interest in making data available for secondary use but articulated concerns about doing so responsibly. Communities working with qualitative data have explored methods for de-identification, and may need further support to implement mediated access to these materials. The benefits of effective preservation and discoverability coupled with appropriate access drive responsible research, supporting scholarship by indicating the availability of data, and creating opportunities for longitudinal research in these areas.

General data collection challenges

(matters that are potentially shared with other domain RDC's.)

As with all research domains, limitations on computational access to data and facilities or resources for sustainable retention impacts the maturity of digital research methods within disciplines. Commitment to computational accessibility and data retention can create opportunities not only to use data within certain kinds of research but can increase opportunities to share data from diverse sources in online platforms or analysis environments, or use data within accessible digital tools.

Discoverability, accessibility, interoperability and reusability of data are key to:

- Improving efficiency of research and making the most of our research resources
- Expanding the scale and opportunities for new research
- Transforming research practices, creating new intra-, inter-, and transdisciplinary methods, and enabling collaboration

Considerable resources are dedicated to the maintenance of data collections in diverse environments. These initiatives preserve and manage a lot of relevant data, providing a rich vein of material for HASS research, but the landscape would benefit from coordination and consolidation, and institutional limitations often limit the availability of data to interfaces and systems that are modelled on traditional HASS research methods. This leads to limited interoperability of platforms, and limited accessibility of data.

RDC-ready collections

The HASS RDC proposal, detailed later in this report, identifies certain HASS research collections demonstrating an advanced state of readiness to participate in and benefit from a HASS Research Data Commons. The projects described below form key elements of the HASS Research Data Commons proposal and are therefore described in more detail.

Linguistics Data Commons of Australia

During 2019 and 2020, a group of linguists has worked to develop a proposal for a Linguistics Data Commons of Australia (LDaCA). Working collaboratively to articulate technical, policy and engagement elements of the Commons, they have iteratively built a five year roadmap for the LDaCA, and worked to develop community endorsement. These processes have informed two short term ARDC Data and Services projects (*Modularising PARADISEC 's catalogue as a model for the data commons*, and *Overcoming pinch-points in ingesting, cataloguing and accessing (meta)data for the development of a national language data commons*). These completed projects have explored challenges in the development of a data commons, including management and implementation of interoperable formats, the technical challenges of ingesting data and metadata about community languages, and making complex linguistic annotations machine readable. These projects are complemented by a wide ranging audit of nationally significant languages data. This coordinated approach across funded opportunities indicates a high degree of community collaboration and a mature approach to engagement with a commons.

Challenges addressed by the Commons proposal

- Strategically important collections of Australian languages data are dispersed, and in some cases, inaccessible and at risk
- There are opportunities to capitalise on existing project-based investment in collecting and researching Australian and regional linguistic data, and elevate digital research approaches through increased coordination, collaboration and accessibility

Key elements

- Identify distributed nationally significant collections
- Build portal to existing data sources, harvesting metadata into an aggregator as a single point of access to the collections
- Develop access policy framework and engagement strategy
- Roll out tested data packaging technologies, increasing sustainability and versatility of linguistics data technologies
- Improved pipelines for data ingest, aiding sustainability of data developed through publicly funded research
- Developing services across collections, including ability to build concordances across multiple collections; wordlists and dictionaries; ability to build assemblages
- All facilities to be developed with Indigenous community input and guidance, recognising the value of data to communities beyond the research environment

Benefits

- Opportunities to strategically utilise Australia's rich linguistic resources to contribute and capitalise on emerging economic opportunities: including speech-to-text technologies, advance development of natural

language processing technologies (integral to maximising commercial use of large, opportunistically collected data), language data used for training algorithms, including internet search mechanisms.

- Provide world-leading outputs including the ability to identify inference from across a range of languages, which has immense real-world applications for Australia. These range from going beyond searches for explicit instances of hate speech when parsing communications to identify people with malicious intent (e.g. terrorist organisations), tracking the spread of information through social networks, through to systematically tapping into inferences that can be drawn from communications when assessing the well-being of individuals in vulnerable communities (e.g. farmers at risk of depression, community response to crisis).
- Australia is one of the most linguistically diverse nations in the world, with researchers exploring Indigenous languages, Australian English(es), community languages in Australia, as well as the languages of the southwestern Pacific. This aligns with national priorities to support support research on Australian indigenous languages, nation building and community, and multilingualism
- Create new opportunities to connect with large-scale international initiatives, including CLARIN (European Research Infrastructure for Language Resources and Technology), advancing international collaboration, contributing to discipline development in Australia and ensuring Australian input on transnational opportunities.

Integrated Research Infrastructure for Social Sciences

While there is considerable investment in social sciences research infrastructure from governments and institutions, including AURIN, PHRN, the Australian Data Archive, ABS Datalab, Multi-Agency Data Integration Project, Centre of Excellence for Children and Families over the Life Course (the Life Course Centre), and many other initiatives, this landscape remains fragmented, which holds back the scale of research, the development of new research methods and frameworks, places limits on collaboration, and introduces inefficiencies to research workflows. This also limits how data-enabled research methods and resources are flowing into researcher training. Each of these initiatives brings considerable expertise and activity that can be leveraged as part of the HASS Research Data Commons.

Challenges addressed by the Integrated Research Infrastructure proposal

- While there has been investment in social sciences research infrastructure in Australia, new forms of research cannot be implemented without better coordination or expansion. Existing research infrastructure is either too small, underdeveloped, or has limited integration with other infrastructure
- There is extraordinary value in supporting social sciences research, which can maximise use and impact of a variety of government-funded data, and lead directly to evidence-based policy making that impacts community wellbeing and productivity
- Rapid expansions in the availability of data point towards new research opportunities, but researchers' ability to capitalise on them is limited by limitations of research infrastructure
- Fragmentation of the social sciences data and infrastructure landscape prevents that community from capitalising on advances in other disciplines working in data-enabled research, including machine learning, artificial intelligence and blockchain
- Much social sciences relevant data is sensitive or has significant privacy considerations. Development of environments and methods for analysis is complex and costly, and beyond the capacity of a single research or data-providing institution. Secure data facilities and support for data integration and linkage can power new research opportunities

Key elements

- Foundational infrastructure for acquisition, storage, documentation and dissemination of social science data, including mechanisms for capture, preservation and analysis of real or near real time data
- Access to quantitative and qualitative social sciences data in a stable, long-term curation environment
- Data linkage and integration support, including secure physical and virtual facilities for enabling access to and analysis of sensitive data. Such support will also benefit humanities research communities calling for further data linkage support.

Benefit

- Integrated Social Sciences Research Infrastructure will capitalise on disparate investments across the social sciences sector, building a cost-effective and accessible data integration and linkage environment
- Such infrastructure will afford social sciences researchers new opportunities to work at scale, in response to emerging datasets that promise rich understandings of our community, including social media, web traffic and the Internet of Things
- Expanded access to data will necessitate access to cloud or high performance computing in social science research, allowing researchers to exploit new data in new ways to inform policy and decision-making

Trove: a platform for data analytic tools

The National Library of Australia's Trove service was raised repeatedly by researchers as an invaluable research resource, indicating that an enormous amount of data-enabled humanities and arts research would be impossible without it. The transformative effect of access to digital resources including digitised Australian newspapers, Picture Australia, Libraries Australia, biographical lists and other informational infrastructures over the past decade is undeniable. Trove's approach to providing large quantities of computationally accessible data, through web services and APIs was ahead of the field of digital initiatives in the Australian GLAM community, with many only attempting similar services very recently. Trove has demonstrated unmatched foresight into the needs of the computational HASS researcher.

In recent years Trove has worked to make 20 years of the .au web domain accessible, aggregated considerable nationally significant data from other GLAM collections, including state libraries and community archives, and has launched the National eDeposit Scheme, facilitating digital deposit of all material published in Australia or by Australians. Trove access services presently struggle to keep pace with the material available and the demands of researchers, while their resources are limited within current funding arrangements.

Challenges addressed by the platform proposal

- Emerging big datasets, including the Australian Web Archive and data collected through the National eDeposit Scheme, remain computationally inaccessible due to restricted resources for interface development
- Trove's existing interface is widely used by a variety of researchers, and graduate students and PhD candidates, who are not engaging with more sophisticated technologies. This proposal would provide capacity to bridge into deep digital techniques, using Trove's popularity within emerging research communities to embed long-lasting cultural change and skill development. It would also, operate as a hub for more advanced research communities to develop tools.
- Would facilitate integration of Trove sourced corpora with analytic tools

Key elements

- A 'researcher portal', accessible through Trove, facilitating access to various data analytic tools, working across a wide variety of data corpuses.

- Tools would include machine learning, and would extend from 'basic' digital tools, through to more advanced usage. In time, advanced users would have opportunity to collaborate on tool development.
- Using a cloud service, the platform could be used to access data from other sources, including access to datasets with restricted access, without data being handled by Trove.
- Platform would be supported by sophisticated accreditation and ongoing helpdesk functions.

Benefit

- Trove occupies a unique position in the HASS research community, with a trusted record of high quality data provision. It outstrips all other efforts computational access to GLAM data, and has demonstrated advanced readiness for the researcher using GLAM sources in non-traditional ways. Recognition of Trove as research infrastructure, beyond the traditional role of the National Library of Australia, will garner considerable support from the research community.
- Facilitating new forms of access to big data including the Australian Web Archive, materials subject to legal deposit and other corpora promises rich new frontiers in study of Australian culture, history and community, areas of high interest to the Australian Government's research agenda, as demonstrated by recent investment in Special Research Initiative for Australian Society, History and Culture.
- Development of a data analytics platform in an established, trusted and well used environment like Trove will benefit cultivation of new researcher skills, new communities and embed long-lasting cultural change.

Proposed foundations, priorities, and goals of a HASS Research Data Commons.

Between September 2019 and January 2020, the ARDC consulted widely across the HASS research and data provider community (details in Appendix 3). This has comprised a mix of individual, small group and large group consultations, as well as participation in relevant community events. The project has also been informed by the ARDC's Open Call for funding for research platforms and services, conducted during October and November 2019, which attracted a large number of HASS-focussed applications. These platform proposals provided additional clarity and emphasis on community priorities and opportunities within HASS fields.

The consultations inform all sections of this report, including characterisation of the HASS landscape, stakeholder considerations and the recommendations proposed. Key priorities that emerged through consultation and have impacted the HASS RDC proposal are outlined below.

From the work of this project and others it is clear that the creation of a single platform to resolve HASS research priorities would be extremely challenging. A portfolio approach, comprising a range of strategic activities able to test and address particular priorities that logically link to and build on each other, has been proposed and there is general agreement that this approach has a greater chance of success and engagement.

Furthermore, access to secure analysis environments, linkage of disparate data collections, and data governance are generic challenges which can be effectively addressed through a portfolio of activities that complement, are informed by, and inform other aligned projects.

The target for this activity is the development of a coordinated and integrated suite of resources and services which enhance the capability of and create new research opportunities for researchers within the broad umbrella of "HASS". While it is important that all of these parts work together to deliver a more powerful whole, delivery of the component activities by identified long-term "owners", or responsible organisations, will assist with the sustainability of those outputs. Organisations leading these activities would be expected to make significant initial investments in the proposed activities and commit to ongoing support, to ensure the outcomes are valuable and valued and are not lost or abandoned.

Improved aggregation of, access to, and sustainability of data sources and collections

Including:

- Government data
- Data sourced from galleries, libraries, archives and museums
- Data generated through research

As discussed earlier in this document, accessible collections drive innovative and high-quality research and collaborations, lend themselves to linkage and aggregation, and make international linkages possible while inaccessible collections result in limited and/or missed research opportunities.

Fragmentation and inconsistency of accessibility in the HASS data landscape has been highlighted as a key inhibitor to the development of new research directions, methods and skills in the community.

During the consultations researchers particularly expressed a desire for improved and better integrated access to government data with a particular emphasis on improved access to administrative data. This priority closely aligns with an imperative to improve linkage across and access to sensitive data collections, discussed in further detail below.

Researchers, data providers and others reiterated the potential benefits of better access to GLAM collections, particularly to nationally significant but as yet un-digitised collections. Several high-impact approaches to maximising the value of GLAM digitised collections and improving their availability to researchers was surfaced during discussions, especially with respect to newly available digital collections.

GLAM sector custodians of valuable data collections have demonstrated high interest and engagement in improving collaboration with researchers to maximise the use, and value, of their collections. Evidence of this can be seen in the widely endorsed Santa Barbara Statement on Collections as Data, and recent efforts to bring researchers, data scientists and collections together in GLAM Lab environments. Successful development of useful aggregation environments could extend to other collections in time, including community held datasets and private sector data.

Researchers expressed concern about missed opportunities caused by inaccessible or lost ARC funded research data in the absence of environments for sharing, and saw great value in data sharing. This can help them fulfil their responsibilities under the *Australian Code for Responsible Conduct of Research (2018)* and can drive new research through re-analysis or linkage for new approaches.

This priority will be guided and enabled by implementations of agreed models for Indigenous data governance and improved take up of best practice in managing and providing access to sensitive data. These are essential considerations in any proposal to increase access to data about people.

A key consideration which has not had significant strategic or coordinated thought revolves around the sustainability of existing and new collections. Many key research data collections across HASS and other domains are subject to uncertain and unpredictable funding support, meaning they are at significant risk. The development of models of long-term sustainability is a key activity of the ARDC and, in this domain, will directly benefit the planning roadmaps of data-generating and holding organisations.

Improved environments for analysis and collaboration

In general, group consultations observed that workbenches or services made with adaptability and reuse of tools as the key priority have, to date, often produced platforms which are so generic that their usefulness is limited. Clearly there will always be a preference for bespoke solutions precisely matching users' wishes and, while myriad custom solutions is not a realistic or desirable outcome, the assertion that quality research will be more effectively addressed through activities responding to more specific user needs is, we believe, compatible with a portfolio approach to solution development, as specific use cases can be leveraged to develop flexible outputs that can subsequently be translated into alternative disciplines.

Our consultations, and the outcomes of the ARDC Platforms program, indicate that a practical approach to providing research platform infrastructure for HASS would be to implement the portfolio approach described above through a set of discipline-specific or research community specific activities based, where possible and practical, on an approach to building interoperability and reuse into the underlying technology, working towards a longer term vision for an agile and flexible national research infrastructure framework spanning not only HASS but other linked domains.

This resonates with successful research infrastructure investments in Europe, where alignments and umbrella infrastructures are under development and in use, clustering mature and tested discipline or research community specific platforms. Initiatives including the Social Sciences and Humanities Open Cloud[12], and DARIAH's Humanities at Scale[13] have developed in this way, aligning discipline and domain specific investment in the first instance, and country specific investment in the second.

Improved data linkage for improved outcomes, better secure analysis environments and agreed models of Indigenous Data Governance

HASS research uses data that concerns people. This means there is a direct link between the outcomes of HASS research and community wellbeing. It can, however, mean that the ethics of collecting, linking, handling, analysing, sharing and reusing that data is subject to or can impose limitations on some forms of research. Tested models of ethics approvals and guidelines for responsible research have created frameworks that guide research practices in all domains, not just HASS, but specific action on certain key issues can support new developments in data-driven HASS research.

Commitment to Indigenous data sovereignty affects the management and access to Indigenous research data, practical implementations of this have remained elusive, which limits community access to data and creates uncertainty for both sector and community researchers. Alignment of key stakeholders in the design and delivery of a practical implementation of Indigenous Data Governance Guidelines will clarify issues for researchers and their collaborating communities.

Research communities and data providers have expressed a desire to test practical models for implementing Indigenous data governance, and this will be a necessary part of any discussion about increased data accessibility.

There are also calls for data linkage services and secure analysis environments for working with sensitive data. It should be noted that the current absence of such environments does not keep sensitive data secure, it either limits access to data for research or actually puts sensitive data at risk as it is dependent on the data handling skills of individual researchers.

Several existing models of location or discipline specific secure analysis environments exist. Solutions for the HASS community can consider the expertise of these services, or build upon existing services as part of this initiative.

[12] About SSHOC

[13] *Design and Sustainability Plan for an Open Humanities Data Platform v 1.1*

HASS Research Landscape - future state

Principles

Future infrastructure should enable highly skilled digital HASS researchers, working in multidisciplinary teams, to use and contribute to the development of the best digital methods to generate new insights into society, culture, and the past while providing reliable and effective forecasting on key social issues, strategies, and outcomes into the future.

HASS researchers should have access to complete, thorough, timely, organised and informed data about today's society and digital access to valuable GLAM collections locally and internationally. Close integration of data and computing services will allow efficient and innovative HASS research into understanding our past and future, providing a sound and trusted base for informing policy and service delivery decisions and working with researchers from other fields to drive benefits for other domains and society. Large datasets will inspire and benefit from new modes of analysis, including technologies such as machine learning, and attract communities of innovative leading researchers and train a pipeline of emerging researchers.

International Context

Research is international and collaboration and integration with international activities and resources has the potential to particularly benefit HASS research through the aggregation of data sets, the ability to leverage sustainability initiatives, and exposure to novel tools and resources.

International models illuminate a path towards new research futures and, with the current potential to strategically invest in a "fast-follow" model, learning from international activities, Australia is ideally placed to benefit from those experiences. Over several decades, coordinated and strategic European investments in particular have resulted in mature discipline-specific and country-specific platforms as well as collaborative tools and the development of rich, collaborative, and productive datasets.

For example, the launch of the Social Sciences and Humanities Open Cloud (SSHOC) in January 2019, a cluster project within the European Open Science Cloud, crystalised the alignment of disciplinary and other more specific investments over preceding periods. SSHOC, to be delivered over nearly four years by 20 partner organisations and 27 further associates, connects mature HASS projects from the European Strategy Forum on Research Infrastructures, including:

- CESSDA: Consortium of European Social Science Data Archives, large-scale, integrated and sustainable data services for Social Sciences
- European Social Survey: cross-national survey measuring attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations
- SHARE: Survey of Health, Ageing and Retirement in Europe, micro data on health, socio-economic status, and social and family networks of more than 120000 people
- CLARIN: Common Language Resources and Technology Infrastructure, digital language resource facility
- DARIAH: Digital Research Infrastructure for the Arts and Humanities, pan-European infrastructure for arts and humanities scholars working with computational methods, and

- E-RIHS: European Research Infrastructure for Heritage Science , supporting heritage interpretation, preservation, documentation and management.

The project ... aims to effect the transition from the current data landscape with its disciplinary silos and separate facilities to an integrated, cloud-based network of leveraged and interconnected data infrastructures. These data infrastructures will be supported by the tools and training which allow scholars and researchers to access, process, analyse, enrich and compare data across the boundaries of individual repositories or institutions[14].

Rather than creating new centralised infrastructure, this initiative connects mature, discipline and subject specific infrastructures that have demonstrable impact for the EU social sciences and humanities.

Other large scale international models that demonstrate the potential of research infrastructure support for HASS include:

- A $1 million Andrew W. Mellon Foundation supported Computing Cultural Heritage in the Cloud (CCHC) project at the US based Library of Congress. The LC Labs team will test a cloud-based approach for interacting with digital collections as data. In collaboration with subject matter experts and IT specialists at the Library, LC Labs will invite a cohort of research experts to experiment with solutions to problems that can only be explored at scale. The Library of Congress announcement of this project explains that "the Library's digital collections comprise a treasure trove of data whose research potential is only beginning to be realized. LC Labs — the Library's digital innovation team — is now looking forward to how the Library, and other cultural heritage institutions, can free huge digital collections for modern computational research"[15].
- A CLARIAH VL project that is using Flemish supercomputing resources to use machine learning and computer vision techniques to automatically assign metadata to scanned objects, images and text. Natural language processing tools are then analysing texts semantically and syntactically, operating as a proof of concept project for automated analysis and annotation[16].
- European Historical Population Samples Network, providing a directory of linked population data initiatives, enabling population reconstitution and life course studies[17].

We therefore propose two core avenues of opportunity to drive towards a proposed "ideal" future state:

- Leveraging and enhancing existing existing NCRIS capabilities (i.e. ARDC, PHRN and AURIN) and other national and institutional resources and activities to underpin a HASS RDC.
- strategically investing in a portfolio of programs under a new NCRIS HASS capability "umbrella" to address challenges which will directly and immediately benefit HASS communities and create solutions which other disciplines and domains can make use of.

[14] About SSHOC
[15] Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data, October 2019
[16] CLARIAH VL Open Humanities Service Infrastructure Work Plan 2019-2020
[17] European Historical Population Samples network

Leveraging and enhancing existing Australian capabilities and activities

Government activities supporting open data:

The concept of a HASS RDC clearly aligns with recent Australian Government policy announcements with respect to the benefits of more open research data, and consultations around the *Data Sharing and Release Act* have been discussed elsewhere in this report. The ARDC engages with government processes and policies related to research data at each opportunity and strongly supports and encourages increased accessibility of data. This report notes specifically that the 'Five Safes' model of mediated data access that inform the proposed Data Sharing Principles in the *Data Sharing and Release Legislative Reforms Discussion Paper* are already in use in relevant areas of social sciences data management and sharing, including the Australian Data Archive, and continued alignment and cooperation with the Office of the Data Commissioner will enable a HASS RDC to achieve maximum access to data sourced from government agencies.

The CSIRO Data61 "Making Australian Government Data Accessible" (MAGDA) initiative, supported by the Department of Treasury and Finance, works with Federal Government agencies to make high value datasets available across government and to the public, while working to ensure appropriate privacy settings are applied. The intended outputs of the MAGDA platform are powerful new data portals that maximise the discoverability and reuse of high-value public data in the government, industry and community sectors. As well as supporting internal agency activities, the open source platform now provides the underpinning capability behind the *data.gov.au* portal. Data61 have also delivered the NationalMap initiative, a map based interface that allows the overlaying of spatially enabled Federal and State Government datasets across a broad range of topics, including agriculture, infrastructure, transport, and social and economic data. The ability to leverage the common interests of the MAGDA, NationalMap and other Data61 activities will ensure a HASS RDC has access to the broadest scope of government, industry, and community data sets.

Finally, efforts to improve researcher access to Government data through the Commons will benefit from capitalising on the Location Index (or Loc-I) project, and other initiatives emerging from the Data Integration Partnership for Australia. The focus within these projects on integrating data on people, business and the environment clearly demonstrates common areas of interest across government and the research community which will benefit both sectors.

NCRIS Capabilities

"NCRIS is a national network of world-class research infrastructure projects that support high-quality research that will drive greater innovation in the Australian research sector and the economy more broadly. Projects support strategically important research through which Australian researchers and their international partners can address key national and global challenges"[18]

The NCRIS is a world-leading research infrastructure program comprised of a portfolio of research infrastructure projects ranging from nuclear, to marine and fisheries, and medical administrative data. As such NCRIS itself

[18] NCRIS Website https://www.education.gov.au/national-collaborative-research-infrastructure-strategy-ncris

provides a rich source of existing and reusable skills and resources which will be key components in the development of a HASS national research infrastructure.

NCRIS capabilities and potential activities include:

Astronomy Australia Ltd (AAL)

AAL facilitates access for Australian-based astronomers to the world's best research infrastructure, encouraging the sharing of astronomical technical capabilities to maximise their value to the nation, and inspiring Australians with these astronomical achievements.

Potential Areas of Common Interest: Image processing, machine learning, high performance computing

Atlas of Living Australia (ALA)

The Atlas of Living Australia (ALA) is a collaborative, digital, open infrastructure that pulls together Australian biodiversity data from multiple sources, making it accessible and reusable.

The ALA helps to create a more detailed picture of Australia's biodiversity for scientists, policy makers, environmental planners and land managers, industry and the general public, and enables them to find, access, combine, and visualise data on Australian plants and animals.

Potential Areas of Common Interest: Data access models, Citizen science, Indigenous knowledge management, sensitive data

Microscopy Australia

Microscopy Australia enables access to an array of high-end microscopy platforms and associated technical expertise in strategic locations to efficiently service Australia's microscopy needs, including optical, electron and X-ray techniques.

Microscopy Australia also has formal connections with a range of other specialised linked laboratories and researchers in fields as diverse as biology, metallurgy, archaeology, engineering, energy and immunology have all benefited from this facility.

Potential Areas of Common Interest: Imaging and Image processing, Image management and sharing, rights management, sensitive data

Australian Research Data Commons

The Australian Research Data Commons (ARDC) is a transformational initiative that enables Australian researchers and the eResearch community access to nationally significant, leading edge data intensive eInfrastructure, platforms, skills and collections of high-quality data.

Potential Areas of Common Interest: FAIR, Data access, management & curation, platforms, policy, community development

Australian Urban Research Infrastructure Network (AURIN)

AURIN enables Australian planners and researchers to make informed decisions about future infrastructure and urban environments based on realistic scenarios and evidence-based analysis.

AURIN has an unequalled, and globally unique ability, to source and integrate a myriad of clean data sets that can provide the 'right answer' to urban planning problems, at all scales of development. In using AURIN, Australian city planners can join landmark 'smart cities' like Singapore, Chicago and Amsterdam as leaders in planning and sustainability.

AURIN is a national collaborative network of leading researchers and data providers across the academic, government, and private sectors. AURIN provides a one-stop online workbench with access to thousands of multidisciplinary datasets, from over 100 different data sources.

Potential Areas of Common Interest: Linkage of social sciences, government and other urban data using spatial tools, analysis environments, building thematic research networks and adaptable interfaces, community engagement, access and interoperability of data from diverse government sources.

Integrated Marine Observing System (IMOS)

IMOS routinely operates a wide range of observing equipment throughout Australia's coastal and open oceans, making all of its data accessible to the marine and climate science community, other stakeholders and users, and international collaborators. This national system for observing the ocean helps us to better understand climate change in Australia and improve our international collaboration and cooperation.

IMOS observations are turned into data that can be discovered, accessed, downloaded, used and reused in perpetuity by their data facility: the Australian Ocean Data Network.

Potential Areas of Common Interest: Data access models, Community development, analysis and annotation of audio visual and image data, management and analysis of sensor or other machine sourced data, image management.

National Computational Infrastructure (NCI)

The National Computational Infrastructure (NCI) is Australia's leading research-sector high-performance data, storage and computing organisation, providing expert services to benefit all domains of science, government and industry.

NCI brings the Australian Government and the Australian research sector together through a broad collaboration involving the largest national science agencies, universities, industry and the Australian Research Council.

Potential Areas of Common Interest: High performance computing, data curation and FAIR

National Imaging Facility (NIF)

The NIF has established a national network that provides state of the art imaging of animals, plants and materials for the Australian research community.

Potential Areas of Common Interest: Imaging and Image processing, Image management and sharing, rights management, sensitive data

Pawsey Supercomputing Centre (Pawsey)

The Pawsey Supercomputing Centre is one of two, Tier-1, High Performance Computing facilities in Australia (the other being NCI), whose primary function is to accelerate scientific research for the benefit of the nation.

This project has built a high performance computer which will prioritise research in geosciences and radio astronomy, and process data including that produced by the Australian Square Kilometre Array Pathfinder radio telescope.

Potential Areas of Common Interest: High performance computing, data curation and FAIR

Population Health Research Network (PHRN)

The Population Health Research Network (PHRN) is a national collaboration that enables the vast amounts of information about Australians, a valuable national resource which can be used to improve our understanding of disease, develop treatments and improve services, to be brought together and made available for important research.

This project provides researchers with the ability to link de-identified population health data from a diverse and rich range of health data sets, across sectors and jurisdictions. This supports nationally and internationally significant population-level research that will improve health and wellbeing and enhance the effectiveness and efficiency of health services.

Potential Areas of Common Interest: Data linkage, collaboration with state and territory data linkage units, sharing sensitive data, development of secure analysis environments

Terrestrial Ecosystem Research Network (TERN)

TERN observes, measures and records critical terrestrial ecosystem parameters and conditions for Australia over time from continental scale to field sites at hundreds of representative locations. This information is standardised, integrated and transformed into model-ready data, enabling researchers to discern and interpret changes in land ecosystems.

This network enables ecosystem scientists to collect, contribute, store, share and integrate data across disciplines. It encourages collaboration and nationally consistent data and promotes understanding of ecosystem change, the rate of change, and underlying causes is essential for effectively protecting and managing Australia's environment and the many services it provides.

Potential Areas of Common Interest: Data access models, spatio-temporal data management, analysis and sharing, management and analysis of sensor or other machine sourced data, sensitive data.

Other national and institutional resources and activities

Time-Layered Cultural Map (TLCMap)

This LIEF supported initiative led by the University of Newcastle is building a tool for mapping data in the humanities, with the added dimension of time. This open and interoperable project is making data linkage possible through geographic elements. It draws on the Australian Heurist initiative for interpreting and sharing humanities data, and connects with the European Recogito tool, which enables collaborative data and document annotation. The project will also connect with other Pelagios supported tools.

The project draws upon the University of Newcastle's Centre for 21st Century Humanities' successful production of the Colonial Frontiers Massacre Map, which received international acclaim for its new models of gathering, analysing and sharing historical data. The Time-Layered Cultural Map brings together eleven chief investigators from diverse disciplines and different universities, leading a range of projects that include visualisation of the geographic and temporal distribution of domestic violence reports, and mapping geo-located records of traditional Indigenous knowledge in Western Australia. The first year of this five year project is supported by the ARC LIEF program, with University of Newcastle committing to the remainder of the project. This initiative supports a collaboration and data linkage environment for researchers working on projects with a spatiotemporal analysis and presentation element. This experimental and presentation environment would complement and be supported by a HASS RDC. The tool can be supported by a HASS RDC where researchers can both find data for linkage and analysis, and share data they have used in their research.

The Digitisation Centre of Western Australia

$1.1 million of LIEF funding was recently announced for a collaborative digitisation facility involving five Western Australian Universities, State Library WA and the WA Museum. This world-class archival quality Digitisation Centre is described as "a major piece of national research infrastructure" and promises that it "will have the capacity to digitise all significant Humanities, Arts and Social Sciences (HASS) research collections held by participating institutions within a decade". This large scale approach to digitisation of HASS collections, with a stated research focus, could be a model for further GLAM digitisation and their intention to digitise all significant HASS research collections will inspire a flush of new research opportunities. A HASS RDC will ensure maximum discoverability and utility of that new data, while this new initiative is a timely testbed for efforts to improve interoperability and discoverability of HASS data.

Linked Semantic Platforms for Policy and Practice

This APO-led LIEF supported project brings together the Analysis and Policy Observatory, Australian Data Archive, AURIN and the Home Modification Clearinghouse, with the intention of using linked open data methods to improve interoperability across major social science databases and develop new analytical tools that improve capabilities for evidence-based policy making. The project establishes interoperability through shared taxonomies, database interoperability for linking policy documents and underlying data, and text mining for references and enhanced metadata. Outcomes of this experimental project demonstrate the potential and value of semantic links in improving interoperability between major Australian data collections, which will inform new approaches to aggregation of similar data sources on a larger scale.

Data Co-operative Platform for Social Impact and Wellbeing

The LIEF supported Data Cooperative (Co-Op) Platform for Social Impact and Wellbeing had its origins in an ARDC Data and Services project during 2019. The platform is developing infrastructure to support data integration and harmonisation of diverse data resources to make data-driven research and decision-making in the social sciences more effective and efficient. This project will tackle the array of data types and sources, to streamline data analysis and drive innovation across a range of critical social issues. It will power research supporting healthcare, better outcomes for disadvantaged and vulnerable groups, resilient urban, rural and regional communities, and increasing our capacity to respond to climate change. Consulted HASS researchers have expressed interest in analysis environments for sensitive data analysis. Existing infrastructure exists for sensitive data analysis, each with limited subject or geographic focus. The HASS RDC should draw from the experience of these initiatives as we consider approaches that would scale for accessibility by more disciplines and researchers.

ARDC Platforms

In December 2019 the ARDC announced a range of investments in research-oriented platforms and services that connect and provide access to a range of resources for researchers and industry. Among the ten platforms supported by this $12 million investment are three HASS focussed investments:

- Coordinated Access for Data, Researchers and Environments (CADRE): CADRE responds to increasing availability of Australian Government data by establishing a shared and distributed sensitive data access management platform for the social sciences and related disciplines to enable data owners and users to address the core concerns around governance, creation, management and sharing of sensitive data for research. The platform will combine a Five Safes (safe People, Projects, Data, Settings and Outputs) implementation framework for sensitive data, with pilots involving secure access environments, cloud storage, data linkage environments and urban data analytics.
- The Australian Transport Research Cloud (ATRC): The ATRC will enable researchers to analyse and model the impact of potential policy changes and planning decisions on transport and travel behaviour, allowing more informed decision making. The platform will provide integrated access to a range of transport specific high quality datasets (ABS Census and Journey to Work datasets, State Household Travel Surveys, National Road network datasets, real time traffic and public transport data, people flow data, public transport timetable data, and de-identified smart travel card datasets) and transport network analysis and simulation tools.
- FAIMS 3.0 Electronic Field Notebooks: Capturing data in the field is expensive, error-prone, and inefficient. FAIMS will produce and manage mobile apps for a range of disciplines including archaeology, ensuring data is captured consistently and synced to the cloud. The platform will use modern web technologies and integrate with existing analysis environments to provide a seamless workflow.
- The EcoCommons Platform will include ongoing support and development of the EcoCloud platform and associated data explorer and microservices. This provides an accessible, lightweight environment for analysis using R and Jupyter Notebooks, integrated with data sourced from CSIRO Knowledge Networks, and has previously been adapted to support HASS researchers in the Tinker Studio. Continued support of this proven platform means that it may be re-badged and deployed to support a range of research types. R and Jupyter notebooks are proving to be the preferred analysis tools for a lot of data-enabled research, including HASS research.

This suite of supported activities will inform and integrate with the HASS RDC as it emerges, and link ARDC with key actors in the research, data provision and research support community.

Community initiatives

The project has identified a number of new and emergent initiatives that will impact methods of accessing and using data in HASS research. They are briefly described below:

NSLA National eDeposit Scheme:

Launched in May 2019 the National and State Libraries Australasia promises a massive uptick in availability of Australian published works in digital form. All publishers are legally required to deposit a copy of their works in state or territory libraries, a practice that has ensured the preservation of works published by Australians and in Australia in our national collections. This includes ephemera, journals, newsletters as well as books and other longer works. Developed over four years, the National eDeposit Scheme, or NED, makes it possible for those works to be lodged digitally. The system is operated by the National Library of Australia, with metadata exchanged back and forth between the National and State and Territory libraries. Integration with the Trove system means that all material that is open will be open through existing full text search and data mining activities. The impact of this project is not to be underestimated, in the six months since launch more than 11,000 resources have been lodged using the NED scheme.

Australian Web Archive

Incorporating the Australian Government Web Archive, and twenty years of content harvested through crawling the .au domain, this immense collection of Australian internet activity is a research game changer. Presently comprising approximately 13 billion files, this big data collection has extraordinary research potential, helping us understand Australian culture, major historic events and how they were described by contemporary participants, the rise and fall of cultural movements, and the paths that have led us to where we are today. The massed digital collections built by Trove are a rich vein of research potential, ready to be exploited as new tools, methods and research communities are supported to make them useful. This and the National eDeposit Scheme, also administered by Trove, are some of the genuine opportunities for big data research in humanities, arts and social sciences, and will be a great basis for seeding the Commons and building researcher capability as we explore ways to unlock more GLAM sourced data.

QUT Digital Observatory

Comprising the Digital Media Observatory, the Australian Music Observatory and TrISMA: Tracking Infrastructure for Social Media Analysis, this infrastructure initiative is at the cutting edge of collecting and analysing Australian use of digital and new media. The Digital Observatory has established technical and organisational infrastructure for tracking, collecting, and making accessible collections of continuous and live digital data that are of interest to researchers independent of their specific topical, thematic, or disciplinary orientation. TrISMA was developed from a LIEF supported collaborative project by Queensland University of Technology, Curtin University, Deakin University, Swinburne University, and the University of Sydney, as well as the National Library of Australia, mapping the Australian twittersphere and tracking and collecting Australian use of social media since 2015. This real time resource allows large scale data analysis to track trends, use of Australian media, and community responses to crises, political events and natural disasters. The Digital Observatory is accessible to partner organisations and accredited researchers. It is an important large dataset that is powering a great deal of cutting edge digital research, including into how automated decision-making is affecting cultural production and consumption based on technologies such as youtube and streaming services, children and parents navigating the

complex issues of young people's use of social media platforms, digital inclusion in rural Australia, and technology-facilitated violence and abuse[19].

CSIRO Knowledge Networks

This initiative of CSIRO works to aggregate research datasets that are published openly. The project uses linked data protocols to support cross dataset queries. Built to tackle the dual challenge of lack of researcher awareness of open data, and difficulties getting that open data into the researcher pipeline, the project started with open data from CSIRO and data.gov.au, providing access and search through APIs, allowing data to be brought directly into R, python and Jupyter Notebooks powered analysis environments. They have continued to add data sources, and their data pipeline now powers the ecocloud and other virtual laboratory environments. This allows researchers to draw data from multiple sources within an analysis environment that is built for their discipline. This example of aggregation of data from multiple sources with varying conditions of interoperability into a discipline-tailored environment will inform development of discipline or community specific research data commons. These targeted commons investments will either inform or comprise a more unified virtual infrastructure in time.

AIATSIS Indigenous Research Exchange

In 2018 AIATSIS received funding from the Department of Prime Minister and Cabinet to develop the Indigenous Research Exchange. In addition to supporting research to support Indigenous self-governance, this initiative is working to establish a platform for exchange of Indigenous knowledge, drawing together data from government, academic and research organisations, and the private sector.

Mukurtu (NSW) and Storylines (WA)

Supported by State and Territory libraries, the Mukurtu and Storylines platforms provide methods of sharing digitised Indigenous archival material in ways that can be governed by communities. Access to sensitive material is limited, while opportunities for annotation or enhancement of records is enabled, to capture, preserve and share community knowledge. These initiatives have broken ground in exploring new models of information management and organisation, and have demonstrated the potential of community-focussed platforms for presentation of data. Mukurtu is an open source content management system that has been developed by Washington State University and has been implemented in the USA, Canada, New Zealand, suggesting that it can aid interoperability and international collaboration around the organisation, sharing and cultural security of information.

Emerging GLAM Labs network

'Labs' are emerging in Australia and internationally as a model for supporting digital interaction with GLAM collections. DX Lab at the State Library of NSW, and University of Newcastle's GLAMx Lab have existed for some time, while South Australia's North Terrace Cultural Precinct Innovation Lab is just getting off the ground. Representatives of these groups have collaborated with international partners including the British Library, Library of Congress, Ghent University, Qatar University and others to cultivate an international community and promote this method of attracting digital experimentation with cultural heritage collections through the production of the book Open a GLAM Lab. The Lab model has been commended as a model for bringing together

[19] Examples of Australian research using this and other digital media platforms showcased at the recent Association of Internet Researchers' conference in Brisbane, October 2019.

researchers, collections staff and data scientists, such collaborations helping overcome gaps in skill, approach or awareness of digital collections to build new research approaches.

Digital Preservation Coalition, Melbourne

In 2020 the international Digital Preservation Coalition has opened an Australian office at the University of Melbourne. The not-for-profit DPC is an international advocate for digital preservation, helping members around the world to deliver resilient long-term access to digital content and services through community engagement, targeted advocacy work, training and workforce development, capacity building, good practice and standards, and through good management and governance. This initiative is evidence of professional interest in preservation and accessibility of growing digital collections in the academic and GLAM sectors, and will build on the success of a grassroots movement, Auspreserves, that has gained much community support in recent years.

Proposed HASS NRI: “The Human Observatory”

The success of a HASS Research infrastructure investment relies on it being broad-based, forward-thinking, innovative, and able to respond to changing challenges. The task is to arrive at solution(s) that enable innovative research today, overcoming contemporary limitations, with an eye to the future. We must think about tomorrow's researchers and what they will need to take advantage of and develop the research approaches of the future, based on and guided by the experiences, opportunities and issues faced by today’s research communities.

The core questions therefore are: “What foundations need to be laid today to build on for success tomorrow and who are the key stakeholders to drive sustainability?”

The ARDC has identified a suite of national infrastructure opportunities that we believe will form a robust and successful foundation for a HASS investment framework. This suite of activities provide immediate and high impact outcomes for the HASS research community, and by extension Australia, and build an effective pathway towards a powerful and productive HASS Research Data Commons. The proposal responds to the key principles identified through consultation with relevant communities and outlines activities that can model the development of a commons over a five year investment, with a pathway toward realisation of a HASS Research Data Commons evident at the five year horizon.

The cluster of activities that make up this proposal respond to the key principles identified through consultation with relevant communities. These focus areas reflect the concerns and ambitions of HASS researchers, data providers and align closely with the relevant Commonwealth Government initiatives:

- **Aggregation of and access to existing data**
- **Improved analysis environments for aggregated data**
- **Data linkage for improved outcomes**
- **Secure linkage and analysis environments**
- **Indigenous Data governance**

It should also be noted that there is potential for a HASS NRI to act as a communication and community development locus and catalyst for cross-pollination across the numerous HASS activities described in the section “Leveraging and enhancing existing capabilities and activities” as well as others. Experience from the ARDC and other national research infrastructures has shown that a highly effective and powerful tool in an NRI’s kit is the ability to gather diverse stakeholders and activities together in a neutral and collaborative space, encouraging free and open discussion which all participants benefit from. It would not be expected that there would be a coordination role, as all of the HASS activities are diverse in funding, participation, and governance, but it is anticipated that outputs from such forums would include identification of potential shared services (such as skills development, user support and/or operations) as well as sharing best-practice and initiating new collaborations. We therefore add one more, non-technical, avenue of activity to the proposed NRI:

- **HASS sector communication and collaboration**

We have grouped these activities under the working title of “*The Human Observatory RDC*” to accentuate the breadth of scope and longitudinal nature of the activities we propose are optimal in enabling HASS researchers

to position themselves, their tools, and their data assets to most effectively access, share, analyse, and draw conclusions from the massive and varied collections of information that inform HASS research.

From the archaeological record, through GLAM collections, to contemporary collections of social data and administrative government data, there are rich veins of material that, given the right tools, HASS researchers can mine to help us understand ourselves and our world. As we face times of significant and rapid change, the coherent, accessible, and agile proposed national HASS research infrastructure will allow us to make the best use of resources to inform decision-making for the future.

The proposal identifies research communities and data providers in an advanced state of readiness to pilot the key Research Data Commons focus areas for targeted initial investment, complemented by specific activities to tackle challenges that can limit or inhibit research, implementing solutions around data governance, public trust and community stakeholders that will be built upon to form an agile and effective research platform framework.

The first five years of development are described and costed in this proposal. These are activities that will lay the foundations for the development of the HASS Research Data Commons and are not the HASS Research Data Commons itself. If we consider something like the European Social Sciences and Humanities Open Cloud as our ideal state, neither the HASS research communities or data providers are in a position to deliver that infrastructure directly.

Outcomes and deliverables of this initial cluster of activities will become apparent within 12 months of initiation, enabling the planning and execution of future development fleshing out the RDC. Expansion into additional disciplines and communities of data providers will be informed by these initial investments.

Beyond the five year horizon, we see researchers working at a national scale contributing to new tools development, and the availability of data snowballing as research becomes more data intensive and sharing becomes normalised. This is the point at which we deliver the HASS Research Data Commons as a critical component of the national research infrastructure landscape.

Timely investment in infrastructure to support data-enabled HASS research will support and strengthen emergent research practices and stimulate collaboration as new methods, data sources and disciplines develop. Without investment to lift Australian HASS data-enabled research capacity, Australia risks the most efficient use of public research funds, and the loss of research leaders to well supported international initiatives. Continued efforts to work at a national and international scale using dispersed, institutionally or short term funded projects perpetuates inefficient practices and missed opportunity.

Project governance

Governance models are not proposed or outlined in this report, but are considered in our supplementary report *Potential Governance Models for a HASS RDC*. We have explored and nominated these high-level governance principles upon which a successful model is likely to rest.

Successful delivery of the HASS Research Data Commons will be guided by and communicated through a range of structures and communication activities. Ideally this will include a high level advisory body, bringing in research leaders, representatives of stakeholder bodies including learned academies, and leaders of other relevant initiatives, including the GLAM and NCRIS Community. The involvement of research leaders in an advisory and governance capacity will lead to many benefits, including:

- Research leaders' contribution to the implementation of platforms and other elements of this proposal will help ensure they are relevant and needed by those communities, and that interfaces and other entry points are accessible to the communities in question. These leaders can also act as Champions, aiding uptake of the infrastructure, and acting as conduit for broader research community input as the infrastructure develops.
- Strategic appointment of research leadership to governance bodies will also ensure a productive development path for later iterations of the HASS Research Data Commons. The proposal described in this report recommends targeted investment *in the first instance* to communities and data providers displaying an advanced state of readiness. Successful implementation of those first stages will lead to demand across other research communities and parts of the data landscape. Having research leaders involved at the highest level will ensure positive engagement with research communities who are preparing for future investment, and will aid the claims made in this proposal that later rounds of investment will learn from the first stages. Forward planning and cross community engagement will support re-use of tools and platforms and aid interoperability.
- Researcher involvement will likely prompt research projects that make use of new infrastructure, both validating approaches and contributing to further development. In several international examples, companion research projects are funded alongside infrastructure development, demonstrating the feasibility of the platform for the benefit of funders and the broader community.
- Ensuring that research leaders are familiar with the developing data-rich landscape will help them guide curriculum development, including introducing new skills that researchers of the future will need.

The role of this advisory body would be to develop and validate priorities and strategy, in alignment with the other national research data commons activities. This will ensure compatibility and integration and allow for leveraging of other activities to drive a HASS research data commons which is coordinated with and able to benefit from the evolving national RDC model.

Social License

While not specifically within the remit of this study, success of a HASS RDC will depend on a recognition of how important it is to change the prevailing culture around access to and use of government data. There is significant sentiment advocating increased community representation in governance (as in, members of the general public having a formal role in governance, much like lay members of ethics committees). It is useful to note that governance models are currently researcher heavy and may benefit from increased Indigenous and society inputs.

Proposed Activities

Theme 1: Data

Objective:	Improved access, re-use and enhancement of more high-quality data for researchers: particularly Government data, data sourced from the Galleries, Libraries, Archive and Museums (GLAM) sector, and other significant research datasets.

Benefits:	• Make extant data as useful as possible
	• Enable efficient and effective researcher access to increasingly available government data
	• Prove the value of researcher access to GLAM collections at scale
	• Enable new international links for Australian GLAM collections and associated HASS research, including:
	  ○ improved access and use of our Pacific and regional collections, linking with stakeholder communities
	  ○ building links to expose Australian collections in relevant international initiatives, ensuring recognition of our cultural and material heritage in global environments

Focus Areas:	• aggregation of and access to existing data • data linkage for improved outcomes

Activity 1.1: Improved access to Government data

Increase and improve research sector access to government data by leveraging the CSIRO Data61 *Making Australian Government Data Accessible* (MAGDA) initiative, Multi-Agency Data Integration Project, Data Integration Partnerships Australia, and other activity by the Office for National Data Commissioner to improve research sector access to administrative and other government data.

Rationale for investment:	• Improved access to administrative government data has been identified as a high priority by social sciences and humanities researchers in consultation • High costs and lengthy delays in the provision of administrative data are impacting research approaches • Opportunities exist around emerging government initiatives around the proposed *Data Sharing and Release Act*, and ensure a researcher perspective in these developments	
Development time:	**Stage 1:** Improved access to open government data; improved federated search **12 months**	**Subsequent activities:** Integration with social sciences data activity, including sensitive data analysis environments **24 to 36 months**

Who would this benefit?	Group 1205 Urban and Regional Planning Division 13 Education (including education systems, curriculum and pedagogy) Division 14 Economics Division 15 Commerce, Management, Tourism and Services (including banking finance and investment, business and management, tourism) Division 16 Studies in Human Society (including demography, human geography, criminology, political sciences, sociology, economics, economic history) Division 18 Law and Legal Studies Division 20 Language, Communication and Culture (including communication and media studies, cultural studies) Division 21 History and Archaeology Division 22 Philosophy and Religious Studies (including Applied ethics, history and philosophy of specific fields)

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Activity 1.2: Trove: a platform for data analytic tools

While TROVE Has been described in this document as a "highly accessible" collection, able to be accessed by researchers, the collection itself and the remit of the collecting organisation (the National Library of Australia) are not designed or targeted for researcher use. In addition there are significant additional collections which could be made available to research.

This project complements but does not replace existing NLA resources, enabling a focus on the delivery of researcher portals accessible through Trove for analysis of a range of data corpora using a variety of tools. Operating as a cloud service, this platform will create tools for visualisation, entity recognition, transcription and geocoding across Trove content and other corpora. This platform will provide basic through to sophisticated research tools and operate as a hub for more advanced research communities to develop their own tools. A leveraged result of this activity will be the potential for translation of these tools into the NLA's public-facing toolset, resulting in increased participation, usage, and impact.

Rationale for investment:

- Trove exhibits an advanced state of readiness among Australian GLAM collections to provide researcher specific services. Demonstrated researcher use of Trove's platform for data analytics can inform further investment in aggregation services for digitised material from Australia's GLAM sector
- Trove is a core component of many of the most successful and innovative digital humanities projects in Australia to date
- The Australian Web Archive and the National eDeposit Scheme represent a hitherto untapped research resource that has the potential to power big data research in the humanities
- Impact analysis of the Atlas of Living Australia has demonstrated significant cultural impact across the collections sector when it comes to open data and accessibility of biological collections, this will drive similar effect in cultural collections

	• Leverage state and territory government investment in digitisation initiatives, and enable re-use of digitisation that occurs as a product of research • Effective delivery of this service will enable integration with relevant international initiatives, including Europeana, Digital NZ and others, enabling transnational research into global contemporary and historical issues.				
Development time:	**Stage 1:** 3 years, with minimum viable product accessible in 15 months		**Subsequent activities:** After 24 months, exploration of further digital aggregation services for the GLAM sector should commence		
Who would this benefit?	Division 16 Studies in Human Society (including demography, human geography, criminology, political sciences, sociology, economics, economic history) Division 19 Studies in Creative Arts and Writing (including professional writing, performing arts and writing, visual arts and crafts) Division 20 Language, Communication and Culture (including communication and media studies, cultural studies, linguistics, literary studies) Division 21 History and Archaeology				
Year 1	Year 2	Year 3	Year 4	Year 5	Total

s 47C

Theme 2: Platforms

Objective: Demonstrate uplift in research practices and outputs through investment in collections, tailored interfaces and analysis tools.

Benefits:
- Model and test options for improved access to research collections and salvage of collections at risk, data linkage and integration with improved analysis environments.
- Leverage and enhance existing social sciences research infrastructure for improved outcomes through increase in scale.

Focus areas:
- Aggregation of and access to existing data
- Improved analysis environments for aggregated data
- Data linkage for improved outcomes
- Secure linkage and analysis environments

Activity 2.1: Linguistics Data Commons of Australia

Development of the Linguistics Data Commons of Australia (LDaCA). The Australian linguistics community has developed a model for capitalising on existing infrastructure, rescuing vulnerable and dispersed collections, and linking with improved analysis environments for new research outcomes.

This activity will serve as an exemplar and demonstrator of the benefit of shared digital research infrastructure.

Rationale for investment:	• LDaCA exemplifies challenges and opportunities common to many HASS communities, as demonstrated through consultation, including
	  ○ the potential of aggregated data from dispersed sources
	  ○ demonstrating value of re-useable research data
	  ○ salvage of at-risk research data
	  ○ using Indigenous language collections and other collections with complex access requirements
	• Industry and policy-development opportunities present themselves through this proposal, including improved data sources for speech to text technology and machine translation, as well as opportunities through inference analysis to track community sentiment
	• Australia's competitive advantage in linguistics research, one quarter of the world's languages are spoken in this region, many of which are well archived
	• Opportunities for linkage with large collections including Trove and the Digital Observatory
	• Benefits for the region, as existing languages collections and researchers work closely with collaborators, archives and stakeholder communities in the Pacific region
	• Opportunities to connect with large scale international infrastructure, including CLARIN
	• Acts as a pathfinder for other discipline specific investments. Structured historical data, including social and economic history, performance history and demography would be fruitful areas for further investment	

Development time:	**Stage 1:** 3 years, with minimum viable product accessible in 18 months	**Subsequent activities:** Pathways to further Commons investment should be considered after 12 months. This is evident in increased support in later years of this proposal.
Who would this benefit?	LDaCA directly benefits Division 20 Language, Communication and Culture Further commons investment will benefit other HASS disciplines.	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Activity 2.2: Integrated Research Infrastructure for Social Sciences

Major development of existing Social Sciences Data Infrastructure, linking and expanding existing initiatives for augmented impact and providing a coordinated governance model for access to data and infrastructure in social sciences and related disciplines; improved capacity to access, preserve and disseminate quantitative and qualitative social science data sources such as surveys, in-depth interviews or the results from experimental trials, in a stable, long-term curation environment; data integration and linkage environments; secure data facilities for accessing sensitive data, and systems and tools for capturing new and emerging real time or near real time data including social media and Internet of Things.

Rationale for investment:	• Leverage and enhance existing investments across the social sciences sector, creating new opportunities to work at scale • Social sciences research leads directly to evidence-based policy making that impacts community wellbeing and productivity • Expanded access to data will lead to new methods and approaches in social sciences research, allowing researchers to exploit data in new ways to inform policy and decision-making	
Development time:	**Stage 1:** 5 years	**Subsequent activities:** This proposal anticipates ongoing support of large scale social sciences digital infrastructure as an integral part of the Australian research and evidence-based policy landscape.
Who would this benefit?	Division 12 Built Environment and Design (including urban and regional planning) Division 13 Education (including education systems, curriculum and pedagogy) Division 14 Economics Division 15 Commerce, Management, Tourism and Services (including banking finance and investment, business and management, tourism) Division 16 Studies in Human Society (including demography, human geography, criminology, political sciences, sociology, economics, economic history) Division 18 Law and Legal Studies Division 20 Language, Communication and Culture (including communication and media studies, cultural studies)	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Activity 2.3: Frameworks for Secure Analysis

Work with existing activities and stakeholders of secure analysis environments to develop a national framework for environments focused on the analysis of sensitive data. This will involve a highly inclusive approach involving participants within and beyond the HASS sector, aiming to create policies, tools, and approaches which enable collaboration on sensitive data across jurisdictions and disciplines.

This activity will leverage other activities within the Human Observatory program including those related to increased data linkage.

Rationale for investment:	• Leverage and enhance existing investments across multiple research sectors, creating new opportunities to work at scale • A national framework for environments which deal with sensitive data will foster innovative research across and within disciplines, leading to improved evidence-based policy making that impacts community wellbeing and productivity	
Development time:	**Stage 1:** 5 years	**Subsequent activities:** This proposal anticipates the development of a national framework for sensitive data analysis. A clear future pathway exists in integrating international compatibility with local capability, further enhancing research and the evidence-based policy landscape.
Who would this benefit?	Division 12 Built Environment and Design (including urban and regional planning) Division 13 Education (including education systems, curriculum and pedagogy) Division 14 Economics Division 15 Commerce, Management, Tourism and Services (including banking finance and investment, business and management, tourism) Division 16 Studies in Human Society (including demography, human geography, criminology, political sciences, sociology, economics, economic history) Division 18 Law and Legal Studies Division 20 Language, Communication and Culture (including communication and media studies, cultural studies)	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Theme 3: Data governance, sovereignty and linkage

Objective:	Focus on challenges that are particular and characteristic of HASS research data, with an interest in ethics and responsible development of data rich environments.
Benefits:	• Inform and complement other elements of this proposal • Develop, test and implement practical responses to key challenges in data-enabled HASS research
Focus areas:	• Improved analysis environments for aggregated data • Data linkage for improved outcomes • Secure linkage and analysis environments • Indigenous Data governance

Activity 3.1: Indigenous Data Governance

Together with stakeholders and existing activities in Indigenous Data Governance (including the Indigenous Data Network and AIATSIS) deliver a roadmap to a robust, community-accepted solution for responsible management and use of Indigenous data collections, to deliver a pilot implementation within twelve months.

Rationale for investment:	• Provide a practical, demonstrable implementation of the principles of Indigenous Data Governance that can inform and guide other elements of this proposal • Improve access to data and research for evidence-based policy, decision-making and service provision in communities of interest	
Development time:	**Stage 1:** Pilot project to develop model for Indigenous Data Governance - **12 months**	**Subsequent activities:** Inform related activities within the HASS RDC, work in integration of services, extend successful model.
Who would this benefit?	Division 13 Education (including education systems, curriculum and pedagogy) Division 14 Economics Division 15 Commerce, Management, Tourism and Services (including banking finance and investment, business and management, tourism) Division 16 Studies in Human Society (including demography, human geography, criminology, political sciences, sociology, economics, economic history) Division 18 Law and Legal Studies Division 20 Language, Communication and Culture (including communication and media studies, cultural studies) Division 21 History and Archaeology	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Activity 3.2: Improved data linkage support and secure analysis environments

Work with stakeholders and existing analysis environment providers to develop the potential of newly released and integrated data that can be linked for innovative research outcomes. This will involve a targeted approach to increased data linkage, learning from and leveraging existing services including state and territory linkage units, PHRN, ABS Multi-Agency Data Integration Project, and the ARDC supported Coordinated Access for Data, Researchers and Environments (CADRE) Platform. This will have considerable cross-fertilisation with Activities 2.1 and 3.1.

Rationale for investment:	• Ensure that the value of increased access to data made possible through Activities 1 and 2 can be realised, by improved linkage opportunities and overcoming challenges in using sensitive data • Extend research opportunities enabled by linked data beyond traditional health and social sciences concerns	
Development time:	**Stage 1:** Refine user needs and identify re-useable components of existing linkage and secure analysis environments, scope development of environment to satisfy unmet need. **12 months**	**Subsequent activities:** Build and integrate scoped environments.
Who would this benefit?	Humanities, Arts, Social Sciences research communities represented by Divisions 12 through 22, inclusive, of the *Australian and New Zealand Standard Research Classification (ANZSRC), 2008, Field of Research Codes.* Developers of the HASS Research Data Commons	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Theme 4: Community communication & collaboration

Objective:	Developing strong, well-connected and collaborative HASS research communities, with an emphasis on communities and groups which are engaged in national or potentially national resource development and/or use.

Benefits:	• identification of potential shared services and resources such as:
	  ○ skills development
	  ○ user support
	  ○ operations and underpinning infrastructures (eg geomapping, AI, etc)
	• sharing and developing best-practices including
	  ○ governance
	  ○ sustainability
	• initiating new collaborations
	  ○ Within sector
	  ○ with Industry & Government

Focus area:	• HASS sector communication and collaboration

Activity 4.1: Community forum 1 - Shared Services & Resources

The ability to share, collaborate and leverage existing skills, resources, and services will be of significant benefit to the HASS community. This activity will provide this access to peers, partners and colleagues for the breadth of the HASS research community through an annual Symposium-style event, highlighting the research enabled by the approaches and outcomes of the HASS RDC and exposing cutting edge data-enabled and data-driven HASS research to the sector and more broadly.

Rationale for investment:	• Involving representatives of HASS research communities and stakeholder bodies will ensure that developments are relevant and needed by their communities, and can aid uptake. • Advocacy for and showcasing of strong and well-connected HASS research communities engaged in national resource development and/or use will illuminate a pathway for less mature communities interested in participating in future developments. • Collaborative relationships with research active members of the community will lead to development of complementary new research proposals demonstrating and augmenting the value of the research infrastructure investment. Communication of these research outcomes will aid public engagement with this investment.

	• Representation from a broad range of HASS communities on these bodies and at events will break down disciplinary silos, aiding communication of best practice and development of multi-disciplinary and transdisciplinary research opportunities.	
Development time:	**Stage 1:** Initial symposium delivered within first twelve months.	**Subsequent activities:** Annual symposium and support of other relevant community events.
Who would this benefit?	Humanities, Arts, Social Sciences research communities represented by Divisions 12 through 22, inclusive, of the *Australian and New Zealand Standard Research Classification (ANZSRC), 2008, Field of Research Codes.* Developers of the HASS Research Data Commons	

Year 1	Year 2	Year 3	Year 4	Year 5	Total
s 47C					

Activity 4.2: Community forum 2 - Governance and Sustainability

Governance and sustainability are critical factors in the success of research infrastructures. This activity will involve future focussed research and other community leaders in an advisory capacity, aiding relevance and impact of the Commons as it develops, and providing a valuable conduit of community input. Such support from the community and for the community will improve likelihood of cross-fertilisation of projects, and integration of HASS RDC into research training and practice.

Rationale for investment:	• Consideration of and planning for effective governance and sustainability will enable the infrastructures outlined in this proposal, as well as HASS infrastructures more broadly, to remain fit-for-purpose and will ensure longevity and impact. • Expert advisory group can act as Champions, aiding uptake of the infrastructure, and acting as conduit for broader research community input as the infrastructure develops. • Ensuring that research leaders are familiar with the developing data-rich landscape will help them guide curriculum development, including introducing new skills that researchers of the future will need.	
Development time:	**Stage 1:** Expert advisory group to be convened immediately, terms of reference to be developed immediately.	**Subsequent activities:** Regular meetings of advisory and governance bodies to continue throughout project, with option to build project and task advisory groups and working parties as needed.
Who would this benefit?	Humanities, Arts, Social Sciences research communities represented by Divisions 12 through 22, inclusive, of the *Australian and New Zealand Standard Research Classification (ANZSRC), 2008, Field of Research Codes.*	

	Developers of the HASS Research Data Commons and collaborators.				
Year 1	Year 2	Year 3	Year 4	Year 5	Total

s 47C

Total Proposed NRI investment over initial five years:

This budget represents the Commonwealth investment into each project (as the sustainability of the RDC rests on the commitment of partners, each activity has proposed co-investment components)

	Year 1	Year 2	Year 3	Year 4	Year 5	Total
Theme 1	s 47C					
Theme 2						
Theme 3						
Theme 4						

Overview of the recommended investment

An overview diagram of the HASS RDC: The Human Observatory proposal is provided at figure 2, overleaf. The proposal describes a cluster of activities that would be delivered in a range of environments, coordinated by an overarching capability. The ARDC has considered and shared a range of governance options for that capability, shared in the supplementary paper *Potential Governance Models for a HASS RDC*.

The HASS RDC capability is largely a blend of digital resources and specialist expertise, integrated in some instances with existing computing and physical facilities. A high level overview of the nature of the investment is provided at figure 1 below. In large part activities will be delivered in collaboration with partners.

High level characteristics of HASS RDC investment

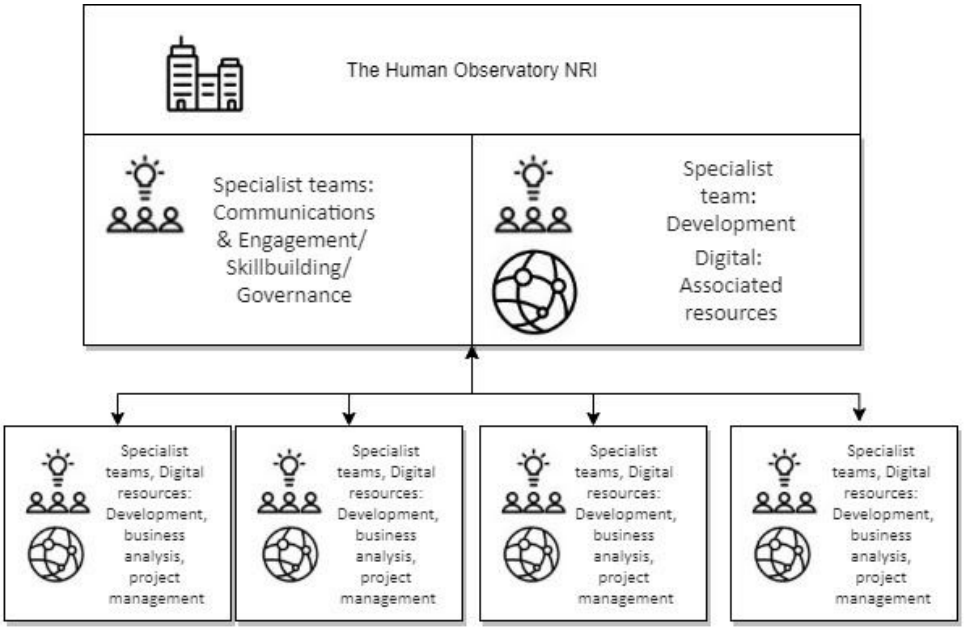

Figure 1: Characteristics of the investment

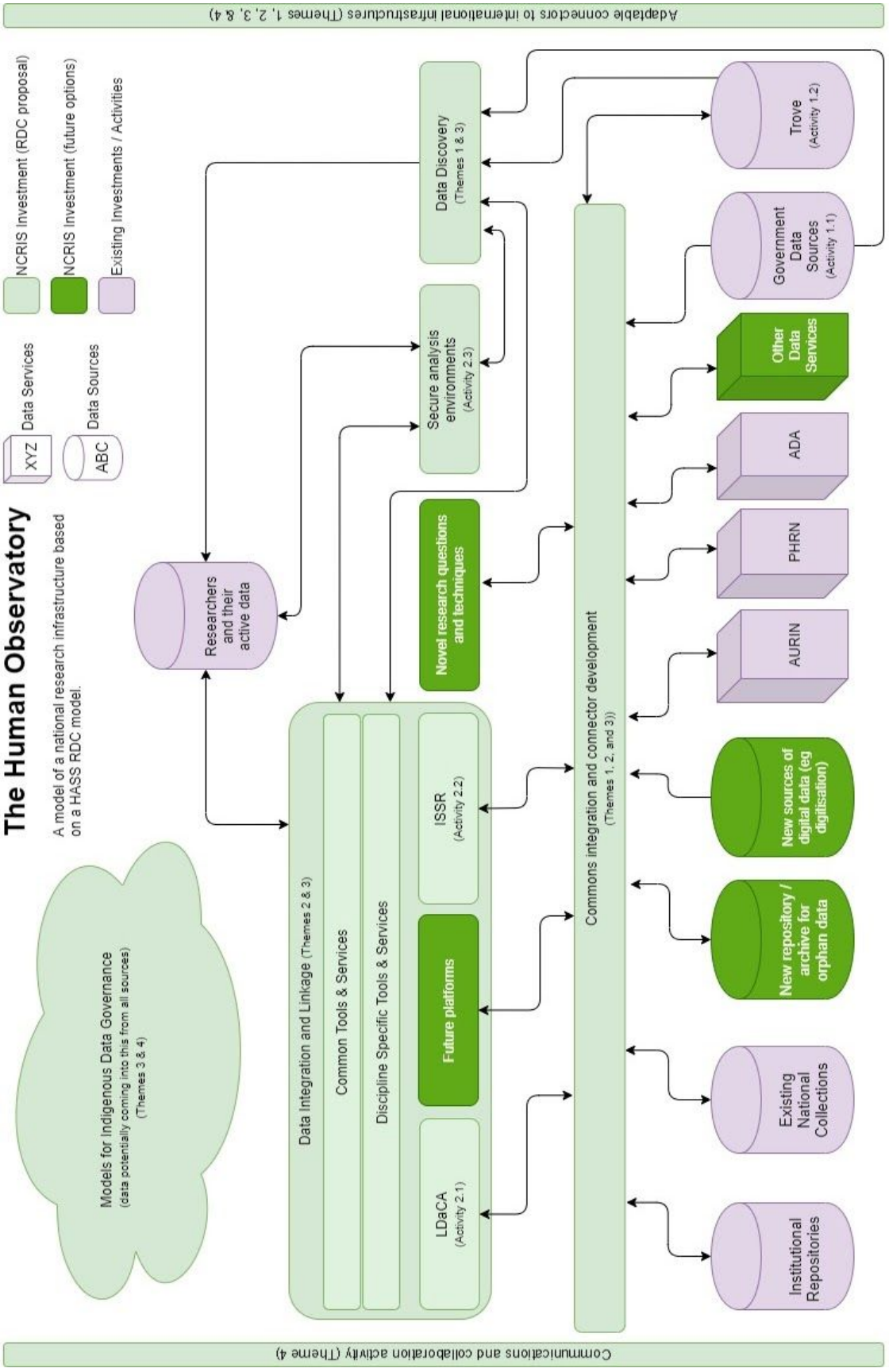

figure 2. Overview diagram

Challenges/Risks

An activity such as the proposed portfolio of HASS RDC projects as components of a HASS will necessarily encompass challenges and risks. The key challenges/risks, impacts, and proposed mitigation strategies are outlined below:

1. Standard software development risks

	Risk	Likelihood	Impact	Mitigation options
1.1	*Inherent schedule flaws*: The HASS RDC proposal sketches a suite of activities, defined to varying degrees of specificity and demanding engagement from a large number of stakeholders and collaborators. The ability to initially tightly define the scope of works is therefore limited.	High	High	• Adoption of consistent project management methodology across all activities. • Contracting of HASS RDC to established NCRIS facility can strengthen underpinning administrative infrastructure • Oversight of HASS RDC by governance body including representatives of stakeholder communities • Simultaneous development of cluster of activities, allowing progress across diverse areas • Ensure sufficient project staff
1.2	*Requirements inflation / Scope creep*: The breadth of activities described in the HASS RDC proposal and the speed of innovation in the digital and data-enabled research environment mean that new opportunities may present themselves, making the project susceptible to scope creep.	Moderate	Moderate	• Adoption of consistent project management methodology across all activities. • Oversight of HASS RDC by governance body including representatives of stakeholder communities • Ensure sufficient project staff • Rigorous review of project planning and timeline • High quality communications activities to drive engagement from cutting edge researchers
1.3	*Key Stakeholder / Employee Turnover*: The cluster of HASS RDC activities will be delivered in collaborative arrangements with stakeholder organisations, creating vulnerabilities related to key stakeholder turnover with	Moderate	Moderate	• High quality communications activities to drive engagement from collaborating organisations, building network of support and commitment • High quality communications activities to drive engagement

	associated risk for the project.			from cutting edge researchers • Rigorous review of project planning and timelines
1.4	*Interoperability challenges*: Drawing data and tools together from a wide range of sources will test the interoperability of those resources. Different disciplines organise and use data in different ways, and the RDC intends to draw on data not initially collected for research purposes. Development of APIs or other connection tools to overcome these hurdles may impact development time frames, while data stewardship resources may need development on the provider and consumer sides.	High	High	• Build contingency into development timelines, accepting that technical components are unlikely to be universally applied and may need adaptation • Investment in training and communities of practice around data stewardship for data providers, government and eResearch support (see also Challenges 3.3 and 3.4)
1.6	*Poor productivity*	Low	Moderate	• Adoption of consistent project management methodology across all activities. • Rigorous review of project planning and timeline • High quality communications activities to drive engagement from collaborating organisations, building network of support • High quality communications activities to drive engagement from cutting edge researchers

The breadth of activity described in the HASS RDC proposal is ambitious and involves considerable stakeholder engagement. At the same time, expectations in the HASS research community and the research infrastructure community are very high.

Mitigation factors described above depend upon cultivating and maintaining a network of support. This includes:

- *Co-location of HASS RDC leadership with an existing NCRIS facility*: creating opportunities to learn from related discipline and domain initiatives and accessing a sturdy administrative foundation.
- *Investment in financial and human resources by collaborators*: Leadership of elements of the HASS RDC can be delivered across a network of stakeholder collaborators, this can cultivate ownership among that community, aiding sustainability, improve pathways for input and ensure broad input at a leadership, governance and strategic level
- *Rigorous project planning, governance and timeline maintenance, and sufficient project staff*: Delivery of simultaneous sub-projects across multiple stakeholders requires rigorous processes of project planning and

review; the project must appropriately budget for staff resources, including for maintenance of project outputs.

- *Communications*: High quality communications can aid stakeholder engagement, cultivate community-wide support as they can observe the project progressing, and can bring in high quality advice, test projects or input.

2. Environment risks

	Risk	Likelihood	Impact	Mitigation options
2.1	*Interdepartmental complications*: Development of some elements of the HASS RDC depends upon distribution of DESE NRI funds to institutions funded and overseen by other sections of government.	Low	High	• Contracting of HASS RDC through established NCRIS facility or other third party can overcome contracting challenges, funding activities on a project basis • Effective communications process to ensure understanding that proposed activities benefit all parties
2.2	*Geographical:* Activity will be dispersed across states and territories and there is a risk of disconnection and poor coordination.	Low	Moderate	• Contracting of HASS RDC through established NCRIS facility allows initiative to work with existing national networks and provides appropriate administrative foundation • Collaboration with PHRN, AURIN and other NCRIS facilities that have overcome the challenge of working across state and territory boundaries can help the HASS RDC benefit from their expertise. These challenges include working with state and territory facilities and managing diverse legal environments around privacy and sensitive data • Oversight of HASS RDC by governance body including representatives of stakeholder communities • Contracting of HASS RDC through established NCRIS facility or other third party can overcome jurisdictional funding challenges, funding activities on a project basis

2.3	*Insecure funding*: The HASS RDC proposal describes a suite of activities to be delivered over five years, with periodic review to respond to a changing research landscape, and a view to a long term and growing investment.	Low	High	• Planning and delivery of the HASS RDC as a suite of components can aid flexibility and responsiveness to an insecure and unpredictable funding environment • Investment of financial and human resources by collaborators, diversifying sources of input, cultivating project ownership and aiding sustainability of elements of the RDC in the case of discontinued funding. • Multi-year funding commitment from DESE
2.4	*Data use and digital skills*: Successful engagement with and maximum impact of the RDC depends upon a broad base of skilled researchers using it, sharing data and contributing to future development.	Moderate	Moderate	• Planning, resourcing and timeline must include skillbuilding and community building opportunities. Activities described under *Theme 4: Community communication and collaboration* are an essential stream of work. • Continuation of existing and effective community engagement activities

Options for delivery of the cluster of activities described in the HASS RDC proposal span organisations that are funded under different branches of government and in different geographical locations, creating complexity in funding arrangements and coordination. Existing NCRIS facilities have demonstrated the value of working with a broad range of partners coordinated by a centralised facility and distribution of project funding by that centralised facility can sidestep the challenges that arise when DESE considers directly funding contributing partners that are normally funded by other segments of government. Doing so also adds clarity, coordination and strong oversight of the outcomes anticipated from this funding stream, rather than topping up existing funding without clear integration with the RDC.

Elements of the future HASS RDC as envisioned depend upon cultural change, growth of new research communities, and embedding digital research methods in mainstream HASS training. This will only be achieved through predictable and stable investment over a number of years. This report recommends in the strongest terms that a medium to long term investment be made under the national research infrastructure program to drive greater innovation and have transformative impact.

Mitigation factors described above spread risk by working with existing NCRIS supported infrastructure, and encouraging community ownership through co-investment, but the Commons will only be realised through coordination as a coherent activity, too wide a dispersion of funds and activities will not deliver the desired outcomes.

3. Sustainability risks

	Risk	Likelihood	Impact	Mitigation options
3.1	*Sustainability of inputs*: The HASS RDC builds on several nationally significant collections and data sources with no ongoing funding	High	High	• Fund activity to enable integration with HASS RDC, aiding understanding of the national significance of contributing collections and seeking alternate and long terms funding sources • High quality communications activities to spotlight collections' involvement, aid their profile and community understanding of their value • Showcase integration with the HASS RDC in new LIEF and other applications by these collections as an indicator of national significance
3.2	*Sustainability of inputs*: poor data stewardship in government leads to lack of access to government data and poor quality data	High	High	• ARDC and community skills programs to increase digital literacy and data curation abilities • Demonstrate increased value of properly curated and managed data for research
3.3	*Sustainability of inputs*: poor data stewardship among HASS community creates low quality data assets	High	High	• ARDC and community skills programs to increase digital literacy and data curation abilities • Demonstrate increased value of properly curated and managed research data
3.4	*Sustainability of inputs*: GLAM digitisation: digitisation of collections from the GLAM sector will provide a very rich source of data for the RDC. Digitisation is however out of scope for this activity.	Moderate	Low	• Demonstrate value of digitisation through use of digital objects in HASS RDC environments • Showcase exposure and use of these collections through the HASS RDC in applications for funding to support digitisation, indicating national significance of these activities
3.5	*Sustainability of outputs:* resources, frameworks, platforms and systems developed in the HASS RDC activity are not sustainable without	Moderate	High	• Ensure strong commitment from participating organisations • Clarity around "ownership" of outputs

	continued DESE funding			• Contracted commitments (co-investments) to future support and operations of outputs
3.5	*Sustainability of outputs*: digital project publication, review, archiving, community data stewardship	Moderate	Low	• Integration of HASS RDC platforms and environments with the ecosystem of persistent identifiers will encourage a changing culture of data sharing and sustainability of digital outputs

Risks explored in the Sustainability table above largely fall outside the scope of the proposed HASS RDC. While integration and collaboration with the HASS RDC may impact the quality of and focus on these activities, the project has little direct ability to address these sustainability risks.

Sustainability of inputs: Consultation informing the development of the HASS RDC has demonstrated that in some cases existing de facto national research infrastructure, including nationally significant data collections and platforms, are probably unsustainable as they do not have access to long term funding. This has an impact on their accessibility and their opportunities to develop and innovate. This proposal has no recommendations for funding for these collections but as measurement of value and impact is a significant focus of the RDC we anticipate the activity assisting those de-facto infrastructures in their search for longer term support and investment.

Similarly, exposure of GLAM collections through the HASS RDC depends upon the increasing availability of more digitised material, which is not funded under the current proposal. The volume of non-digital data in the form of physical GLAM collections remains a persistent and significant challenge for HASS researchers. The ability to progress with a digitisation agenda across HASS disciplines, leveraging the data curation capabilities outlined above, will develop and promote a rich and valuable pool of national data assets. The ARDC expects that demonstrated use of these data through the HASS RDC will add weight to organisations' efforts to seek funding or re-prioritise resources to this activity.

Community data stewardship skills: HASS RDC activity is predicated on the principle that high quality research is and can be driven by high quality data. It is imperative therefore that the researcher communities across the HASS domains are equipped with the knowledge and tools to be able to effectively structure their data inputs and outputs in such ways as to minimise secondary curation effort.

Data stewardship skills and resources in government: There is broad understanding of the value of greater availability of government data among HASS researchers and in government. Sources including individual consultations with the social sciences community, and the experience of Data61 in their work on the MAGDA initiative demonstrated that limited data stewardship skills or resources inside government departments can be an inhibitor to accessing government data. A high degree of risk aversion was described, and in many cases appropriate, but where working relationships had been struck between researchers and suppliers of government data, it was largely in a bilateral agreement. Such directly negotiated relationships cannot be scaled to provide the data sharing environment that both sides agree is desirable. Improved data stewardship arrangements in government are desirable for several reasons, including

- more efficient and reliable decision-making around opening and sharing government data
- more sustainable management and curation of government data
- better preparation of data for archiving and research use
- facilitating authoritative input on development and review of integrated data platforms and secure analysis environments.

Alignment with Commonwealth Government policy objectives and priorities, including response to 2016 Roadmap and Investment Plan.

The development of a HASS Research Data Commons will augment the impact of related policy initiatives of the Australian Government and its agencies. Government funded research and government itself are two key data sources for HASS research. Pressure for greater openness and data re-use are evident in both of those spheres

Responding to the challenges proposed by the key research priorities creates a strong foundation for improved data-driven decision making by government with the research sector able to collaborate effectively and transparently with government and industry in novel and innovative research.

Access to Government data

The 2017 Productivity Commission report into *Data availability and Use*, and the Australian Government's response, encouraged a proactive approach to the release of non-sensitive government data and urged action on methods of sharing sensitive government data. The Productivity Commission's report notes the opportunity for improved data access and use to enable new products and services, improve efficiency, transform everyday life, and allow better decision making, and urges Government not to miss these opportunities due to aversion to risk. It describes the opportunities made possible by release of high value datasets as "rich and long" (Productivity Commission, p. 287), and highlights that "many innovative uses of data cannot be valued right now, as they are not yet envisioned" (p. 284). Similarly, the Australian Government's Public Data Policy Statement (2015), highlights that "publishing, linking and sharing data can create opportunities that neither government nor business can currently envisage", and goes on to commit the Australian Government to "optimis[ing] the use and re-use of public data; to release non-sensitive data as open by default; and to collaborate with the private and research sectors to extend the value of public data for the benefit of the Australian public". In their report the Productivity Commission specifically notes the value of administrative data (as collected through administration of government policies or programs) because it is comprehensive, directly sourced and usually has a high incentive for accuracy.

In their response to this report, the Australian Government committed to, and has delivered, a new National Data Commissioner, and the development of a *Data Sharing and Release Act*. They also agreed with the identification of high value datasets that could "generate substantial benefits across a broad swathe of the Australian population" if shared across and between sectors and jurisdictions (p.291, quoted in Australian Government response to Productivity Commission report).

That response highlights the work associated with Australia's Second Open Government National Action Plan 2018-20, which has also informed the development of the *Data Sharing and Release Act*.

Considerable progress has been made by the Digital Transformation Agency, Data61's *Making Australian Government Data Accessible* initiative and the Multi-Agency Data Integration Project, especially in streamlining access to government data across departments and making certain data available for research. The benefit has not been evenly spread across HASS research sectors, and a greater focus on research needs will lead to broader uptake of these emerging opportunities.

Access to research data

In many consultations with HASS researchers concern was expressed about 'lost' publicly funded research data. Certain cases where 'lost' data was saved were recounted as cause for celebration, including, as an example, where an approach to a retired researcher led to the re-discovery of many 40+ year old recordings that linked to a contemporary project, greatly enriching the resulting research outputs. Identification and preservation of such datasets motivates PARADISEC's 'Lost and Found' project, where they actively seek out researcher recordings of at-risk languages, knowing that there are many valuable research outputs that are under-documented.

At the same time, researchers shared numerous examples of poor data management, many owning up to keeping data from previous projects only on a 'hard drive in the bottom drawer' or the like.

Release of the *Australian Code for the Responsible Conduct of Research* in 2018, and supplementary guide *Management of Data and Information in Research* highlight the principles of transparency and rigour, pointing out that these principles compel researchers to share and communicate data and methodology openly, and ensure good stewardship of publicly funded assets. Failure to adhere to this Code can have real consequences for researchers and institutions, especially threatening opportunities for future funding.

The ARC has also responded to this known issue by compelling improved data management practices as part of funding agreements under the National Competitive Grants program from 2020. Successful grant applicants will be required to "outline how data will be collected, formatted, described, stored and shared throughout, and beyond, the project lifecycle"[20], inline with their responsibilities under the 2018 Australian Code for the Responsible Conduct of Research. Their new contract templates describes this data management plan requirement, expressing the importance of long-term preservation of data arising from ARC funded projects, and comments "We strongly encourage that data arising from the project is deposited in an appropriate publicly accessible discipline and/or institutional repository" (A2.2.5 (c), ARC Discovery Contract Template) This is a welcome initiative that may lead to increased repository demand and changes in researcher work practices.

[20] Research Data Management, Australian Research Council

Appendices

Appendix 1: The question of *National Significance*:

Development of the HASS Research Data Commons will be helped by a definition of what HASS data should be considered "nationally significant".

A robust definition of national significance informs and guides the allocation of resources towards activities which will result in the greatest positive impact in HASS research and data. Landscape mapping has demonstrated that certain de facto national data infrastructure currently depends on rolling short term funding, particularly from the ARC LIEF program (AustLII, AusStage), but also from Centre of Excellence funds (PARADISEC), or on institutional support, which may not guarantee sustainability. Determining a framework of significance will aid the sustainability of this community of HASS data providers, as it would deliver a means of demonstrating their value.

In building a proposed definition of national significance for HASS data this project has considered relevant Australian, New Zealand and international models for determining significance or national interest and/or appraising data or other national assets. Each of these frameworks highlights the role of significance assessment in the allocation of resources to the asset in question.

ARDC National Data Assets

The ARDC is presently implementing a National Data Assets program, which will help us establish a portfolio of national scale data assets that support leading-edge research, through strategic partnerships.

Assumptions underpinning that program are that collections of data can be national research infrastructure when they:

- support leading edge research (research excellence, impact, priorities)
- are national in scale (multi-organisational aggregation, use, and governance)

Productivity Commission report into Data Availability and Use

The 2017 Productivity Commission report into Data Availability and Use outlined a model of National Interest Datasets. Under their recommendations designation as a National Interest Dataset would bring resources for curation and updating, aggregation of jurisdictionally separate datasets and provide a framework in which to better link datasets across sectors and fields of endeavor. Such designation would signify value, but the process is designed around 'additionality'. The report asserts that "this is not about labelling a collection as important in principle, its purpose must be additional national benefit in practice" (Productivity Commission 2017, p. 25, emphasis added)

The Productivity Commission proposes a process of nomination and public review, managed by Parliamentary Committee, agreeing that the notion of 'national interest' can be arbitrary. They do, however, offer characteristics of high value and quality.

The Productivity Commission report distilled input from a range of submissions to determine the distinct characteristics that indicate high dataset value. They

- are unique
- are of high quality
- have a high degree of coverage in the relevant population
- are up to date or updated regularly.

The report indicates that the value of the dataset could be either because it enables wider innovative and beneficial uses or because the dataset enables other datasets to function more effectively (examples of this include the Geocoded National Address File and the Australian Statistical Geography Standard). Additionally datasets should have an established focus on a nationally significant subject matter.

Benefits of designation as NIDs include maintenance as a national asset for at least ten years, listing of all NIDs and managing Accredited Release Authorities (ARA) on a site such as data.gov.au and the report highlights the opportunity for ARAs to curate and manage datasets that do not achieve NID status under the same legislative arrangements.

Data collected through the administration of Commonwealth Government policies or programs, or 'administrative data', is nominated as being particularly valuable because if is comprehensive, directly sourced and usually with an incentive for accuracy.

The Government's response to this report indicated broad agreement with the need for designation of high value datasets, indicating an intention to work on a framework to identify "those datasets whose availability and use will generate significant community-wide benefit" (Commonwealth Government, 2018, p. 9). This should complement existing work under the Open Government Partnership National Action Plan 2016-18.

Digital Curation Centre UK: *How to Appraise and Select Research Data*

Finally, the UK's Digital Curation Centre has developed guidance for the selection and appraisal of research data. Like other significance assessment frameworks the articulated rationale is focused on allocation of resources, but also raises the issue that data becomes more difficult to discover when everything is kept.

The key criteria recommended in that guide are:

- Relevance to Mission: Does it fit the organization's priorities stated in current strategy, including any legal requirement to retain the data beyond immediate use.
- Scientific or Historical Value: Is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
- Uniqueness: The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
- Potential for Redistribution: reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects issues are addressed.
- Non-Replicability: It would not be feasible to replicate the data/resource or doing so would not be financially viable.
- Economic Case: Costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.

- Full Documentation: the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation and use.

Related initiatives in New Zealand

Nationally Significant Collections and Databases Program

The HASS Research Data Commons project brief asks us to give specific consideration to the work of the New Zealand Government in relation to this issue. We have consulted with representatives of the New Zealand Ministry of Business, Innovation and Employment (MBIE) about their Nationally Significant Collections and Databases program. This initiative that has historically supported the management and enhancement of 25 collections and databases determined to be significant. This program is presently under review, the model and level of funding not having been reviewed since 1996, and representatives of the Ministry acknowledge problems with it in its current form.

Key points from the document informing that review:

- Nationally Significant Collections and Databases are closely tied with New Zealand's Statement of Science Priorities, which limits extension to HASS relevant data
- Current criteria 1 & 3 test the collection or database's relationship with strategic support for science from the New Zealand Government, but Criterion 2 queries the national importance of the collection or database against a group of sub-criteria:
    - 2.1: Does the asset make a substantial contribution to the goals set out in the Statement of Science Priorities?
    - 2.2: Is the asset important to a wide range of stakeholders?
    - 2.3: Does the asset deliver substantial benefits to users?
    - 2.4: Is the asset unique nationally and/or internationally?
    - 2.5: Is the asset irreplaceable?
- Current NCSDs were funded on the basis that
    - they are being held on behalf of New Zealand, where continued provision, maintenance and utilisation are critical for New Zealand science to deliver public benefit
    - the benefits accrue to many, varied users and third party beneficiaries while the costs of provision belong to the custodian. (p.11, Scientific Collections and Databases Review: Update Report)
- The review highlights the direct link between long term funding commitments and the capacity for curators to maintain accessible collections and databases.

At the time of reporting the Review has been completed and advice has been prepared for the relevant Minister but not yet delivered. MBIE staff are unable to share recommendations arising from that review until that next phase.

Stats NZ Tier 1 Statistics

In 2005 a set of official statistics were identified as performance measures of New Zealand as a result of a Review of the Official Statistics System. That review determined that "Tier 1 statistics: are essential to central government decision making; are of high public interest; meet public expectations of impartiality and statistical

quality; require long-term continuity of the data; provide international comparability in a global environment." (Official Statistics System, 2007).

Designation of Tier 1 statistics enables prioritisation and rationalisation of statistical investment and effort across the official statistics system.

The relevance of the Tier 1 list is managed through five yearly review. Designation can apply to a single statistic or a set of statistics, and the key characteristics are described as being "relevant, authoritative and trustworthy, provide long-term continuity of statistical information, and enable international comparability" (ibid).

Models of significance assessment in physical collections and heritage

UK Designation Scheme

This program is designed to identify and allocate resources to the protection of cultural collections that are held in non-national museums, libraries and galleries in the UK. The definition of a Designated collection is "A nationally significant, coherent assemblage of items; held in trust in the long-term for public benefit. A Designated collection is an essential research resource for its subject" (Arts Council England, 2015, p.11).

Resources under the Designation program are allocated on a collection basis, not for an organization. A single organization might be custodian of several designated collections. This has parallels in the management of research data collections. This program has allocated £32 million to support 140 collections between 1999 and 2016. These funds are awarded from proceeds of the National Lottery. Designation is an enduring award.

Collections applying for Designated status must demonstrate

1. National Significance
2. Outstanding Quality
3. Research value.

Applicant guidance gives further definition of national significance as 'the collection is focused on a subject of national significance that has had a recognised and lasting impact on society' (ibid, p.17). Applicants are encouraged to consider relevance to episodes in history, geographical or economic development which had profound national impact, evolution of particular communities or schools of thought, or the development of a scientific or artistic movement or form.

Explanation of research value suggests that "the collection is, or has the potential to be, an essential research collection for its subject; the collection makes, or will make, a major contribution to the public understanding of the subject" (ibid, p.19). Further exploration of this encourages applicants to consider whether the collection is "central to advancing public understanding and scholarly knowledge of the subject that it represents", "an established or developing reputation as a research resource" and connection to comparator collections (p. 19).

A 2016 Review re-affirmed the focus of the program – to celebrate and safeguard vital collections for present and future generations.

Significance and the *Burra Charter*

Significance: A Guide for Assessing the Significance of Collections was initially created by the Heritage Collections Council in 2001, and revised and updated by the Collections Council of Australia in 2009. The revised report indicates that users of the method report improved decision-making about collections in areas including preservation, digital access and funding.

The model suggests assessment of collections' significance across four primary criteria:

- historic
- artistic or aesthetic
- scientific or research potential
- social or spiritual

and four comparative criteria:

- provenance
- rarity or representativeness
- condition or completeness
- interpretive capacity

All criteria are considered in relation to all items or collections, but not all will be relevant to every item or collection. The development of significance criteria arose from the recognition across the GLAM community that a framework for decision-making was required in order to make the best use of limited resources.

Conclusion

All considered resources make a strong case for significance frameworks in order to determine responsible allocation of resources. This can refer simply to the storage and management of data, but can also impact decisions by funders and institutions to support research using that data, and can help collections define their value for funders and other decision-makers.

Common themes across the frameworks considered include:

- Being unique, irreplaceable or not easily replicable
- Value to a wide range of stakeholders
- Potential to deliver benefit to many users
- Being comprehensive or having a high degree of coverage of relevant subject
- Are useful, in terms of being appropriately documented, provenanced and ethically governed.

Appendix 2: Definitions

In any document, specific words and phrases can have particular meanings and impact. This is particularly true of research infrastructures and is potentially compounded by the variety and richness of interpretations in the very broad arena which makes up the "HASS & Indigenous" categorisation of research fields. For simplicity and in the aim of ensuring minimal confusion and minimal need for discussion on nomenclature, for the purposes of this document the following meanings are used:

Humanities, Arts and Social Sciences (HASS)

Humanities, Arts and Social Sciences (HASS) covers an extremely broad range of disciplines, a matter that has confounded consideration of research infrastructure investment for this sector in the past. For the purposes of this project, ARDC has defined HASS as disciplines represented by Divisions 12 through 22, inclusive, as defined in the Australian and New Zealand Standard Research Classification (ANZSRC), 2008, Field of Research Codes.

In order to take a practical approach within the limited scope of the project, our exploration has been weighted towards those groups that are highly engaged in the data-enabled research community, and that are closely linked to data providers in the national HASS-relevant data landscape. Similarly, in the interests of practicality, research communities have been considered at varying degrees of granularity. For example, their use of similar data in similar ways, from similar sources, makes it sensible to cluster qualitative and quantitative social sciences as broad groups, while the specific interests of economic historians and social historians has led us to break 'Group 2103 Historical Studies' into different groups. Similarly, the particular needs of researchers working with social media data prompts a specific response, drawing them out of the general 'Group 2001 Communication and Media Studies' group. Finally, the emerging field of Creative Industries has been considered, which does not presently have a FOR code in the Standard Research Classification.

GLAM

An acronym in use by the galleries, libraries, archives and museums community. In some cases this is extended to GLAMR to include records keeping communities. "As the principal repositories of Australia's unique history, art, heritage and audio-visual collections, GLAM organisations have a central role in connecting Australians with the stories and histories of their communities at a local, regional and national level"[21]. GLAM collections have always underpinned an extraordinary range of HASS research, and many GLAM organisations conduct their own research. Digitisation initiatives are offering new research and research translation opportunities.

Sensitive data

Sensitive data are data that can be used to identify an individual, species, object, or location that introduces a risk of discrimination, harm, or unwanted attention. Major, familiar categories of sensitive data are:

- personal data
- health and medical data

[21] From the GLAM Peak submission to the Australian Infrastructure Audit, October 2019

- ecological data that may place vulnerable species at risk

Other categories of sensitive data include legally privileged, commercial-in-confidence, financial records, or data that is culturally sensitive. Culturally sensitive data includes Indigenous knowledge that is restricted to certain communities or community members, or that can threaten sites or practices of significance if widely circulated. HASS data concerns people in many cases, so triggers certain ethical considerations when it comes to storing, preserving and sharing data.

Sensitive data can be published, particularly through the use of the FAIR framework (which does not require data to be open for it to be FAIR). For example, researchers can publish metadata, making data discoverable, without making the data itself openly accessible, which enables conditions of access to be placed around the data. Models of mediating access to sensitive data include the Five Safes model, the Office for National Data Commissioner's proposed Data Sharing Principles, de-identification, and generalisation of data.

Research Data

Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyse the data may also be included[22].

The benefit of digital tools and research methods is that computational analysis can accelerate research or make it more expansive, explore new potential in scale. Researchers' ability to engage in computational analysis depends on having relevant data available and useful. In some cases this will mean creating new data, in some cases it will mean accessing existing data. Increased access to and utility of data can lead to the development of new analysis tools, and of new forms of research, previously unheard of or impractical in a largely analogue research environment.

In contrast to some research domains, humanities, arts and social sciences research data is extraordinarily heterogenous. Traditionally humanities and arts scholars have drawn on archives, library and museum collections, text, survey data, published and unpublished manuscripts, historic observations of people, places, events, works of art, oral histories. Many of these sources are now available in digital form. Social sciences researchers have made use of quantitative and qualitative survey, observational, government, spatial and administrative data. The availability of these data in digital form, and the explosion of new forms of data have created new opportunities in social science, expanded research environments and posed new challenges. New forms of data are emerging in this landscape, including social media, 3D models, sensor data and complex digital objects. The use of digital tools and platforms are creating new interdisciplinary and transdisciplinary applications and new research methods.

Approaches to data in the humanities, arts and social sciences community are more complex than in some domains, given that in most cases, the data concerns people, individuals and communities. Some data is sensitive, for ethical, privacy or cultural reasons. There are unique challenges in the preservation and sharing of

[22] From University of Melbourne 'Policy on the Management of Research Data and Records' as quoted in the ANDS guide What is research data?

personal, culturally sensitive data or qualitative data, which affects how we apply frameworks and protocols for best practice in data use and management.

Research Data Commons

A digital research data commons collocates data, storage and computing infrastructure with core services to enable researchers to conduct and collaborate on world class data-intensive research. As well as enabling access to data, and methods of sharing, a commons can include compute resources, analytical tools and working environments, storage, models and other support.

Research Infrastructure

Assets, facilities and services that support research. This can include physical infrastructure, data, physical and digital tools. Physical and digital infrastructure is accompanied by governance, personnel and support to ensure that infrastructure is effectively and efficiently used for maximum benefit.

Analysis Environments

In this report this general term is used to refer to a range of tools in use by researchers, including online tools and web based environments for research collaboration. Examples of these include the AURIN workbench and the Atlas of Living Australia Spatial Portal. It can also refer to environments that support computational analysis, such as R (an open source programming environment for statistical computing and data analysis), Jupyter notebooks (an open source web application that allows researchers to create shareable documents that include text and code for analysis) and others.

There are various reasons for establishing analysis environments that are beyond the standard desktop or institutional network. These include:

- Efficient and effective access to data from diverse sources, packaged in a way that is discipline or domain-relevant
- Online environments enable collaboration across small and large distributed teams
- Processing high volumes of data is more efficient using cloud computing, rather than desktop resources, and sometimes requires integration with high performance computing facilities
- Working in analysis environments that are collocated with data sources aids responsible and reliable research, as opposed to downloading datasets for analysis. It can ensure that most current data is being used, and reduces risk of researcher error or corruption of data through handling

Collection

For the purposes of this report the concept of a collection means an aggregation of digital resources which has meaning in a research context. This context includes the research process itself, any resources which support that process, and the linked scholarly communications cycle with its research outputs of publications, software and data. Objects from these collections provide context and meaning for each other.

A collection:

- should be understood as a single aggregation of resources within its research context;

- is not comprised exclusively of documents as the output of research, although they can certainly be documents as the subject matter of research; and
- has Australian relevance, either through involvement of Australian researchers, or Australian subject matter.

Data-driven / Data-enabled (hypothesis driven) research

Discussions around research using data use various terms to describe that research, including digital methods, digital research, data-supported, data-driven and the like. In this report the ARDC makes an important distinction between data-enabled and data-driven research, which is linked to emerging trends in research practice across domains.

Data-enabled research is research using data. There are various models of this, but largely this describes traditional forms of research that are augmented, enhanced or where hypotheses are tested against data. The traditional process involves the development of a hypothesis, refinement of a scenario, and then testing that against data, collected or generated.

Data-driven research describes emerging approaches to research that start with the data. This type of research can start with a review of relevant data sources, gathering of those data, analysis of that data (often using algorithmic or machine-learning methods), then developing research questions based on observations and correlations that were found.

Machine-readable

Machine-readable data is data which can be read and interpreted by a computer program without human involvement. In order to be machine-readable, data should be structured in a simple and consistent open format that permits easy interrogation by computer code and does not require the purchase of a specific piece of software or operating system in order to access.

Interoperability

Interoperability refers to the suitability for integration with other datasets, and can refer to the ability to interoperate with applications, related systems or workflows for analysis, storage and processing. Typically a dataset or platform is interoperable if it uses community agreed formats, language and vocabularies.

Many of the data providers consulted in the course of this project asserted that their platforms and data were interoperable. This demonstrates wide understanding of the value of community agreed vocabularies, standard metadata schema and interoperable platforms. Interoperability can depend upon a suitable policy and resourcing environment, however, and can pose challenges when tested.

Data linkage

Data linkage is a method of bringing information from different sources together about the same individual, family, place or event to create a new, richer dataset. Linking information from disparate information sources makes it possible to assess chronological sequences of events, or relationships between inputs, decisions, including policy, and outcomes. Data linkage is commonly used by the Population Health Research Network, and the network of state and territory based data linkage units to provide valuable information for policy and research into the health and wellbeing of the population. Linkage is a mediated process that involves input from

the data custodian, the linkage service or data linkage unit, and the researcher. In current population health applications, linkage units work to separate personal identifying information from service or clinical data. This creates opportunities to work with data from whole populations rather than small sample sets. Linkage between administrative and research or clinical datasets provides an evidence base for policy makers and researchers to better understand population health and wellbeing and implement and evaluate service delivery and programs.

Data linkage is being delivered for further forms of research through the ABS Multi Agency Data Integration Project, while international examples, including Nordic population register projects, demonstrate applications outside population health.

Provenance

Data provenance is the documentation of where a piece of data comes from and the processes and methodology by which it was produced. In the GLAM community provenance documents the history of an object or artwork, recording when it was first made, how it was collected or traded, movements across international borders or from institution to institution, and actions that have been done to it throughout that history. In digital libraries it is used to document a digital object's lifecycle.

In many cases data users are not the producers of their research data. Data producers may configure an instrument or simulation in a certain way to collect primary data, or apply certain methodologies and processes to extract, transform and analyse input data to produce an output data product. Understanding the circumstances under which the data was generated, and processes that have been applied to it is an important indicator of quality for data users.

Before use the data may have been annotated, edited or enhanced by another user, leading to a need to indicate that a new version of the data has been produced, and whose input should be acknowledged.

Providing provenance metadata as part of the published data is important for determining the quality, the amount of trust one can place on the results, the data version, the reproducibility of results and reusability of the data.

In a situation where data is aggregated by a service or third party, high quality metadata indicating provenance can help identify duplicated data, or untangle issues that may emerge where a dataset has been annotated, improved or enhanced and then re-shared, which can create issues around versioning.

Big data, small data

There is no universally agreed 'size' that turns small data into big data, and the characteristics of these descriptors have been debated at length over the last decade or more. Research and commercial applications generally agree on the four 'v's of big data : volume, velocity, variety and veracity[23]. Certain HASS research communities work with data of high volume or veracity, including analysis of social media data, sensor data that informs urban environments research, or, in some cases, corpus linguistics. HASS researchers in many cases observe the value of small data and 'rich data', often referring to the research opportunities created through

[23] As described in this IBM sourced infographic: https://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

linkage of diverse datasets. The variety and velocity of data are a well-understood challenge for data-driven and data-enabled HASS researchers.

Domain/ Discipline/ Community

Domain: Refers collectively to the breadth of HASS disciplines.

Discipline: A more specific group defined by their research interest. Using the ANZSRC Field of Research Codes as a model, it may refer to the two digit 'Division' level, or the four digit 'group' level. See the definition for HASS for more information.

Community: Refers to an organised group within either a domain or discipline. This can be a professional association, or an informally organised group that has formed to deliver a certain activity.

FAIR Data Principles

FAIR applies to a set of principles that were designed by a diverse group of stakeholders at a 2015 workshop in Leiden, representing academia, industry, funding agencies and scholarly publishers. This group developed a set of concise and measurable principles to address how to best enable data re-use. They have since been recognised by decision-makers and funding bodies globally as a model of thinking about data in a way that will enable maximum use and re-use.

The principles are intended to act as a guideline for those wishing to enhance the reusability of their data holdings. They put specific emphasis on enhancing the ability of machines to automatically find and use data, in addition to supporting its reuse by individuals. This makes research data harvestable, increasing the scale and variety of data available for new and innovative processes.

The principles were written in a way to ensure they can be applied across disciplines and across technologies. The principles address both the data and metadata, and recognise that making data FAIR can depend not just on the state of the data, but also on underlying infrastructure, procedures and governance. The FAIRness of data does not depend upon it being open and in many domains data can be kept under mediated access controls and still be FAIR. Humanities, Arts and Social Sciences disciplines have debated the intersection of ethics, Indigenous data sovereignty and FAIR in recent years, which has had an impact on open data advocacy, but the principles remain sound in and use across domains globally.

FAIR Principles (as documented on GO FAIR)

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

Appendix 3: Overview of Australian HASS-relevant collections

The HASS data landscape is rich and diverse. As described elsewhere in this report, researchers draw upon government data, data sourced from galleries, libraries, museums and archives, data produced through research including qualitative and quantitative data. HASS researchers are increasingly working with new forms of data, including social media, web analytics, data generated by sensors, laser scanners and the like.

Comprehensive mapping of this landscape would be a considerable challenge, and significant collections would almost certainly be overlooked. This appendix provides:

- broad observations on sections of the data landscape
- description of a representative group of nationally significant HASS data collections to help inform this report.

The collections described in detail below are not comprehensive, but are representative of the large data collections that power the HASS research landscape.

Institutional repositories:

In the best case scenarios data that is the product of research is deposited in an institutional or other public repository. Most public research organisations maintain an institutional repository with varying degrees of accessibility. Deposit of data into those repositories is governed by institutional and ethics policies, support for depositors provided by the repository or the researcher's institution, and compliance of the data creator with those frameworks.

In many cases nationally significant collections are held in institutional repositories. This includes data generated through research, donations of data from past projects, digitisation or other enhancement of data that is held in institutional libraries, archives or special collections and countless other scenarios. A few examples of nationally significant collections in institutional repositories include:

- The Australian Women's Register, and the Australian Trade Union Archives, managed by the University of Melbourne's eResearch Scholarship Centre
- The Media Archives Project, managed by the Macquarie University Centre for Media History
- Western Sydney University's Women's Oral History Project
- The Australian National Corpus, managed by Griffith University
- Australian Politics and Election Archive, managed at the University of Western Australia

The accessibility of institutional repositories is mixed. Analysis of Research Data Australia, ARDC's data discovery service, including links to many institutional data repositories, showed that across all data less than 6% of datasets were machine actionable (i.e. accessible for direct download), with none of those being the records flagged as HASS relevant. Improved machine readability of data will enable researchers to work with larger aggregations of data for computational analysis, or can aid discoverability of smaller datasets in a crowded and fragmented data environment.

Collections sourced from the galleries, libraries, archives and museum (GLAM) sector:

Digital accessibility of GLAM collections is highly variable. Researchers consulted throughout this project repeatedly identified the lack of digitised collections as a barrier to research, and argued for digitisation initiatives as part of the development of a HASS Research Data Commons.

Collections managed by the GLAM community have formed research infrastructure for humanities, arts and social sciences research since the Great Library of Alexandria. These collections store and preserve not just our expressions of creativity in works of art and literature, our historical records in museums, but also our records of migration and settlement, industry and policy implementation in government and business archives, our approaches to health and human services and our media outputs in public records and film and sound archives. Our historic census records, including population living and service provision arrangements prior to the establishment of the Commonwealth provide us with the longest view on Australians' way of life.

Many state and federal museums, galleries, libraries and archives are actively digitising their collections, powering new research opportunities, that help us understand Australian society, history and culture, as well as the history and culture of our region. This activity is unevenly spread, however, and computational accessibility of the resulting data is mixed, adding to the fragmentation of this research landscape.

Elsewhere in this report, the ARDC has described the advanced state of readiness of the National Library of Australia's Trove platform, while organisations such as Museum Victoria, with API access to their Collections Online facility, are exemplars in this environment. Other organisations provide limited accessibility to their digitised collections. The reasons for this are mixed, including limited resources for the development of digital platforms, and the priority that is placed on digital access to collections for researchers, weighed against access for other users.

Historian and hacker Tim Sherratt maintains an up to date list of computationally accessible GLAM collections that power his GLAM Workbench site.

With their aggregation of data from other digital sources, world-leading Australian newspapers digitisation project, multiple APIs and other forms of computational access, Trove at the National Library of Australia is a standout in this environment. This is the product of collaboration by national and state libraries and is a globally unparalleled research resource.

Privately held archives pose a particular challenge for research. They have no mandate to be publicly accessible, and their sustainability is rarely guaranteed. The Macquarie University led Media Archives Project documents collections of cultural heritage that are in private hands, and have shared concerns about their sustainability. Similarly, during consultation for this project collaborators have shared concerns about tape collections held by community media organisations, especially remote and regional radio and television organisations.

Examples of Nationally Significant HASS Data Collections

Analysis and Policy Observatory (APO)	
Description:	A digital repository of grey literature (material published by organisations whose primary purpose is not publishing) relevant to the full breadth of public policy. Comprising more than 43,000 records in a repository and curated collections, APO is a trusted resource for the policy and research community. Typical material includes government submissions, research reports, products of inquiries, briefings, data, guides and the like. Materials are published by research institutions, government, not for profit organisations, advocacy bodies and similar organisations.
Governance/ Management	APO is a multi-institutional collaborative platform supported by major partners Swinburne University of Technology and the Australia and New Zealand School of Government (ANZSOG). APO is overseen by an Advisory Board consisting of representatives from the major partners, including Swinburne University, RMIT University, UniSA, Australian Data Archive (ANU), Australian Urban Research Information Network, UNSW, Western Sydney University and ANZSOG.
Access and interoperability	The APO metadata approach is aligned with information discovery and organisational needs of both the research and policy communities. The APO-MAP supports metadata creation for bibliographic resources, datasets and research projects. The APO-MAP draws on a range of ontologies fundamental to policy and research management to define metadata elements, including Australian Government Locator Service (AGLS), researchgraph.org, rif-cs (ANDS) and scholix.org. APO also maintains and uses a number of controlled vocabularies, including the APO Public Policy Taxonomy (in development), which is aligned with the scope and structure of the Fields of Research (FOR) and Australian Government Interactive Functions Thesaurus (AGIFT). APO also manages an extended resource type vocabulary that describes the form and/or genre of bibliographic resources. Where possible, APO taxonomies align with those managed by the Confederation of Open Access Repositories (COAR). The current APO led LIEF project - Linked Semantic Platforms - uses AI text mining to develop the APO Public Policy Taxonomy based on subject terms contributed by users, and the database was successfully integrated within an international ResearchGraph trial. The APO metadata approach ensures interoperability with other knowledge graph type initiatives and can be plugged into platforms to enhance connectivity between data sources. APO resources are shared with Informit, which distributes records into university library discovery systems. APO resources are also harvested by Trove and Google Scholar. APO is an approved minter of DOIs (Digital Object Identifiers), providing publishers with a persistent link to their resource, in line with best practice principles in digital publishing.

Users, research communities, other stakeholders	GROUP 1205 URBAN AND REGIONAL PLANNING GROUP 1301 EDUCATION SYSTEMS GROUP 1302 CURRICULUM AND PEDAGOGY DIVISION 14 ECONOMICS GROUP 1503 BUSINESS AND MANAGEMENT GROUP 1603 DEMOGRAPHY GROUP 1605 POLICY AND ADMINISTRATION GROUP 1606 POLITICAL SCIENCE GROUP 1608 SOCIOLOGY GROUP 1801 LAW GROUP 2001 COMMUNICATION AND MEDIA STUDIES GROUP 2103 HISTORICAL STUDIES GROUP 2201 APPLIED ETHICS GROUP 2202 HISTORY AND PHILOSOPHY OF SPECIFIC FIELDS GROUP 1117 PUBLIC HEALTH AND HEALTH SERVICES APO is also heavily used by public sector researchers, not for profit organisations and other researchers outside the academy. Parliamentary Librarians are particularly enthusiastic users.
Other comments	This is an award winning, unrivalled collection of resources that inform policy and decision-making, and can, in many cases, provide more current policy information than formally published research material, as publication is more immediate.

AusStage	
Description:	AusStage provides an accessible online resource for researching live performance in Australia. Researchers and students use AusStage to develop new knowledge about live performance in Australia and to assess the contribution that live events make to the nation's cultural vitality and international image. Companies, artists and reviewers use AusStage to find out who's doing what in the live performance industry. Librarians, archivists and museum staff use AusStage as a source of information on items in their collections and to assist the public with enquiries. Researchers in universities, performing arts organisations and government agencies are working with AusStage to map patterns of live performance over time and space, to explore networks of artistic creativity in the performing arts, and to create opportunities for audience research and development.
Governance/ Management	Flinders University is lead institution as part of a management committee comprising researchers from University of Newcastle, University of Melbourne, Monash, University of Queensland. Advisory council includes state and territory representatives and an Indigenous representative.

Access and interoperability	Data is harvested by several aggregation services including Trove, HuNI, Austlit, and allows programmatic access using a variety of web services, described here. Data model is described and API access is provided to certain geographic and networking functions AusStage Resource Directory draws in resources from external providers (including datasets, digitised materials etc) and is built upon the Dublin Core metadata schema. Identifiers in AusStage are custom and persistent,they presently maintain more than 500,000 URIs.
Users, research communities, other stakeholders	GROUP 1904 PERFORMING ARTS AND CREATIVE WRITING; GROUP 2005 LITERARY STUDIES; GROUP 2103 HISTORICAL STUDIES; GROUP 2102 CURATORIAL AND RELATED STUDIES In use by formal networks of performing arts and drama studies associations. Industry use is also evident.
Other comments	Funded to date (~20 years) through 6x LIEF grants and bridging funds from Flinders University, including co-contribution by partners. Storage and server provision all an in-kind contribution from Flinders University. Ongoing arrangement is that LIEF income funds research, with each co-contributing institution receiving back 1.5x their contribution. In some cases some of that research money goes towards development, done at Flinders.

Australian Legal Information Institute (AustLII)	
Description:	AustLII provides free internet access to Australasian legal materials. The facility publishes public legal information, including primary legal materials (legislation, treaties and decisions of courts and tribunals); and secondary legal materials created by public bodies for purposes of public access (law reform and royal commission reports, commentaries and summaries on the law, and a substantial collection of law journals.
Governance/ Management	AustLII is a jointly operated "research infrastructure facility" of the Faculties of Law at the University of Technology, Sydney and the University of New South Wales (UNSW).
Access and interoperability	AustLII operates as an informational infrastructure in itself, and interoperability is applied selectively across the collections. Copyright of legislation is largely held by crown bodies, so access to full text isn't computationally supported, but the AustLII hosted material is designed in such a away that it can integrate with other legal information infrastructures. Thus AustLII is built on the same frameworks as international equivalents including NZLII and services for the Pacific Islands. The AustLII team are highly interested in improving internal interoperability, which would facilitate machine learning analysis of legal text, review and decision-making, and can open up opportunities for new research. Some sections of AustLII have complex restrictions on access, including their case law records, which are restricted from indexing by google and other search functions, to protect privacy of individuals named in the records. AustLII collaborates with Trove, working on improving and expanding data sharing arrangements, especially around records of government gazettes.

Users, research communities, other stakeholders	DIVISION 18 LAW AND LEGAL STUDIES GROUP 1205 URBAN AND REGIONAL PLANNING GROUP 1301 EDUCATION SYSTEMS GROUP 1501 ACCOUNTING, AUDITING AND ACCOUNTABILITY GROUP 1502 BANKING, FINANCE AND INVESTMENT GROUP 1503 BUSINESS AND MANAGEMENT GROUP 1504 COMMERCIAL SERVICES GROUP 1602 CRIMINOLOGY GROUP 1603 DEMOGRAPHY GROUP 1604 HUMAN GEOGRAPHY GROUP 1605 POLICY AND ADMINISTRATION GROUP 1606 POLITICAL SCIENCE GROUP 1607 SOCIAL WORK GROUP 1608 SOCIOLOGY GROUP 2103 HISTORICAL STUDIES AustLII is also used by the public, the legal profession and government. The source of 35% of users cannot be identified so figures are inexact, but 40% of AustLII users appear to be from the legal profession, 30% from government and 30% are academic researchers.
Other comments	AustLII participates in the Free Access to Law Movement, and plays an important role in a network of legal information infrastructures. They cooperate to ensure international interoperability, and back up international LII services, ensuring ongoing provision of access to legal information on the grounds of access to justice. AustLII has benefited from successive rounds of LIEF funding over 25 years, working effectively at planning and executing relevant extensions to capacity while supporting core functions. The initiative is also supposed by the legal fraternity and other contributors through the AustLII Foundation. The organisation supports the base level of activity with income from the Foundation, while project grants are dedicated to delivering new functions.

Australian Data Archive (ADA)	
Description:	The Australian Data Archive (ADA) is a Core Trust Seal certified repository, based in the ANU Centre for Social Research and Methods (CSRM) at the Australian National University (ANU). ADA was established 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. It provides a comprehensive social science data collection, with a catalogue of more than 6000 datasets from more than 1500 projects and studies from 1838 through until the present day. This includes survey data, opinion polls, censuses and includes international collections from the Asia Pacific region. The ADA acquires, documents, preserves and disseminates data online to a broad range of social science researchers in the university, government and other sectors, and also locates and manages access to international social science data sets sought by Australian based researchers.

Governance/ Management	The ADA is based in the ANU Centre for Social Research and Methods (CSRM) at the Australian National University (ANU), and is supported by team of professional data archivists, advised by a panel of social scientists. Supporting infrastructure is provided by the National Computational Infrastructure (NCI).
Access and interoperability	The ADA is managed on an instance of Dataverse, an open source data repository platform developed by the Institute for Quantitative Social Science at Harvard University in use by many relevant international repositories. The Archive adopts, develops and applies standards in line with international best practice (such as DDI and OAIS model), and is active in the *International Federation of Data Organizations* and the *International Association of Social Science Information Service and Technology*. The ADA works collaboratively on a number of national and international interoperability projects. Accessibility of ADA data is mixed, with open access to some collections, special access requirements on others and highly restricted access on some. ADA's commitment to open access wherever possible is balanced against their obligations to the original participants in research studies. Access is in many cases determined by the ethics commitments of the depositor. The ADA is committed to the 'Five Safes' principle of access to sensitive data, a model that is internationally adopted by many similar organisations, and is recommended in recent material from the Office for the National Data Commissioner. The ADA has participated in many projects aimed at increasing interoperability between platforms, including Humanities Arts and Social Sciences Data Enhanced Virtual Laboratory, the APO led Linked Semantic Platforms LIEF project, and the recently announced Coordinated Access for Data, Researchers and Environments (CADRE) initiative supported by the ARDC.
Users, research communities, other stakeholders	DIVISION 12 BUILT ENVIRONMENT AND DESIGN GROUP 1205 URBAN AND REGIONAL PLANNING GROUP 1299 OTHER BUILT ENVIRONMENT AND DESIGN DIVISION 13 EDUCATION DIVISION 14 ECONOMICS DIVISION 15 COMMERCE, MANAGEMENT, TOURISM AND SERVICES DIVISION 16 STUDIES IN HUMAN SOCIETY GROUP 1603 DEMOGRAPHY GROUP 1604 HUMAN GEOGRAPHY GROUP 1605 POLICY AND ADMINISTRATION GROUP 1606 POLITICAL SCIENCE GROUP 1607 SOCIAL WORK GROUP 1608 SOCIOLOGY GROUP 1699 OTHER STUDIES IN HUMAN SOCIETY DIVISION 18 LAW AND LEGAL STUDIES DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE GROUP 2001 COMMUNICATION AND MEDIA STUDIES GROUP 2002 CULTURAL STUDIES
Other comments	ADA plays a central role in the Australian social science data environment, managing significant sets of longitudinal studies, social attitudes surveys, health studies, election

	studies and opinion polls. The ADA also houses many datasets generated through ARC Discovery and Linkage projects.

AustLit	
Description:	AustLit is an information resource and research environment for Australian literary, print and narrative cultural material. It includes bibliographic history and, where possible, links to the full text of creative Australian literature, including fiction, drama, poetry, children's and young adult literature, travel writing, autobiography, memoir, biography, essays, Indigenous life stories and oral history. The collection includes one million records and 87,000 full text items, as well as useful directories, bibliographics and other informational infrastructures. AustLit also covers critical material on Australian literary works, creative writers and critics, on Australian literature in general and biographical material about Australian writers and other significant figures in Australian literature. Material is also included about organisations concerned with the development and production of Australian literature and its distribution such as publishers, distributors, literary agencies, magazines, journals and newspapers, writers' groups, writer's festivals and the Literature Board of the Australia Council. AustLit also documents information about Australian awards, prizes and literature funding. The repository provides online access to aggregated web journals and texts. Material largely dates from European settlement in Australia onwards (i.e. late eighteenth century), but does include some relevant texts published prior to that. Research using AustLit includes comparative study of literature sourced from different states and territories, and different periods, exploration of responses to crisis or global events as reflected in Australian cultural material, development in the publishing, print and creative industries, and Indigenous cultural expression and its reception. Resources including the Australian Drama Archive, Blackwords, anthologies of criticism, colonial magazines and the like provide a rich text collection ready for research in new ways.
Governance/ Management	Established as a LIEF supported project from 2000 to 2010 the AustLit resource is now housed within the University of Queensland and sustained through partner contributions, contributions for development coming from research grants, and a subscription service.
Access and interoperability	The subscription model required to aid financial sustainability of Austlit limits its interoperability as it is unable to be harvested by Trove or allow any full text export. This limits computational analysis and use of this resource at scale. Data curation and research analysis is presently highly manual. Bibliographic data is organised according to the FRBR standard, which has the potential to be highly interoperable but resourcing and limitations on open access make greater interoperability impractical.
Users, research communities,	DIVISION 19 STUDIES IN CREATIVE ARTS AND WRITING GROUP 1901 ART THEORY AND CRITICISM

other stakeholders	GROUP 1902 FILM, TELEVISION AND DIGITAL MEDIA GROUP 1903 JOURNALISM AND PROFESSIONAL WRITING GROUP 1904 PERFORMING ARTS AND CREATIVE WRITING GROUP 1905 VISUAL ARTS AND CRAFTS GROUP 1999 OTHER STUDIES IN CREATIVE ARTS AND WRITING DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE GROUP 2001 COMMUNICATION AND MEDIA STUDIES GROUP 2002 CULTURAL STUDIES GROUP 2003 LANGUAGE STUDIES GROUP 2004 LINGUISTICS GROUP 2005 LITERARY STUDIES GROUP 2099 OTHER LANGUAGE, COMMUNICATION AND CULTURE GROUP 2102 CURATORIAL AND RELATED STUDIES GROUP 2103 HISTORICAL STUDIES GROUP 1302 CURRICULUM AND PEDAGOGY GROUP 1303 SPECIALIST STUDIES IN EDUCATION AustLit is heavily used in primary and secondary education, and by state, territory and public library services.
Other comments	While the subscription model of AustLit limits interoperability, the resource is functionally accessible according to its original mission, with most researchers, students, librarians and teachers having access through library and school organisational subscriptions. Sustainability is the primary concern of AustLit. Operating costs are not fully covered by subscriptions and the small team of staff spend considerable time contributing to research grant applications.

Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	
Description:	A facility for digital conservation and access for endangered language and cultural materials from the Pacific region, defined broadly to include Oceania and East and Southeast Asia. This is a region of rich linguistic diversity, with over 2000 of the world's 6000 languages spoken in Australia, the South Pacific Islands and Southeast Asia. The collections now holds material from all over the world and more than 1200 languages are represented. PARADISEC has an established framework for accessioning, cataloguing and digitising audio, text and visual material, and preserving digital copies. The digital archive conforms to international archiving standards, and collection practices have a strong focus on the safe preservation of material that would otherwise be lost, especially field tapes, including recordings from the 1950s and 1960s, which were at high risk of loss through degradation. In addition to archiving, PARADISEC is a centre of training in linguistic and musicological data management and linkage, developing data models and integration with online tools.

	Access for interested communities is a high priority, and considerable cultural renewal work is done in partnership with represented communities and cultures. The archive works to make field recordings accessible to those who are recorded and their descendants, and local communities and researchers receive training in language documentation. PARADISEC supports archives in the Pacific region, including the Vanuatu Cultural Centre, the University of New Caledonia, the Solomon Islands National Museum, the Institute of Papua New Guinea Studies and Rapa Nui.
Governance/ Management	Led by a consortium of University of Sydney, University of Melbourne and Australian National University, directed by a steering committee representing those three universities and two directors.
Access and interoperability	Digital outputs are available in various formats depending on users' needs. The collection is catalogued using the Dublin Core metadata standards, and the recommendations of the Open Languages Archives Community, meaning that metadata can be harvested into international aggregators. The International Association of Sound and Audiovisual Archives (IASA) cites PARADISEC as a model of good practice for digital mass storage systems. Open access is preferred and each item in the collection has access conditions specified by the depositor at the time of entry into the collection. PARADISEC encourages the use of Creative Commons licences.
Users, research communities, other stakeholders	GROUP 1601 ANTHROPOLOGY 1904 PERFORMING ARTS AND CREATIVE WRITING (including musicology and ethnomusicology, Aboriginal and Torres Strait Islander performing arts, Pacific peoples performing arts, etc) DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE GROUP 2002 CULTURAL STUDIES GROUP 2003 LANGUAGE STUDIES GROUP 2004 LINGUISTICS GROUP 2099 OTHER LANGUAGE, COMMUNICATION AND CULTURE PARADISEC works closely with international collaborators in the Asia Pacific region, including supporting establishment of local archives and training programs.
Other comments	The international value of this archive has been recognised by inscription onto the UNESCO Memory of the World, the European Data Seal of Approval, a special commendation in the 2016 UK Digital Preservation Awards, a University of Melbourne award for excellence in team-based research and in 2019 they received the international Core Trust Seal based on the DSA-WDS Core Trustworthy Data Repositories Requirements.

Design and Art Australia Online	
Description:	Design and Art Australia Online (DAAO) is a collaborative e-Research tool built upon the foundations of the Dictionary of Australian Artists Online. DAAO is an open source freely

	accessible scholarly e-Research tool that links biographical data about Australian artists, designers, craftspeople and curators with information about artworks, event histories and collection details. Building on initial datasets created through the LIEF supported Dictionary of Australian Artists Online, recent developments of DAAO have focussed on interoperability with related platforms, including Trove and AusStage, and exposing relevant data to international environments, including Wikipedia. DAAO has built project workspace capability to support researcher collaboration, and has opened access for interested parties to contribute via crowdsourcing tools.
Governance/ Management	Managed by University of New South Wales Art and Design, the DAAO depends upon a volunteer board of editorial advisors, a small team of development staff, and engagement from researchers leading related projects.
Access and interoperability	As resourcing for DAAO has scaled back in more recent years, the initiative has effectively implemented a high degree of interoperability, demonstrating an interest in sustainability of data through sharing and collaboration. Their last LIEF project (2014) was focussed on boosting interoperability measures, including expanding their OAI Provider, which enables harvesting through Trove, enriching metadata that supports person and group entities, enabling visualisation in secondary environments, and expanding their RIF-CS feed. Examples of this interoperability include the ability of the Australian Directory of Electronic Literature and Text-based Art (ADELTA), ARC funded infrastructure based at the University of Western Sydney, built an API within the DAAO server environment to use DAAO as a naming service, further embedding DAAO resources into other digital research infrastructure. DAAO is the second largest contributor of person records to Trove's bibliographic resources, and has worked on data sharing projects with partner organisations including the National Library of Australia, Art Gallery of WA, Art Gallery of SA, Queen Victoria Museum and Art Gallery, Australian Experimental Art Foundation, Mildura Arts Centre, Art Gallery of Ballarat, Geelong Gallery, BUDA House, Castlemaine Art Gallery &Historical Museum and Benalla Art Gallery.
Users, research communities, other stakeholders	DIVISION 19 STUDIES IN CREATIVE ARTS AND WRITING GROUP 1901 ART THEORY AND CRITICISM GROUP 1902 FILM, TELEVISION AND DIGITAL MEDIA GROUP 1903 JOURNALISM AND PROFESSIONAL WRITING GROUP 1905 VISUAL ARTS AND CRAFTS GROUP 1999 OTHER STUDIES IN CREATIVE ARTS AND WRITING DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE GROUP 2001 COMMUNICATION AND MEDIA STUDIES GROUP 2002 CULTURAL STUDIES GROUP 2005 LITERARY STUDIES GROUP 2102 CURATORIAL AND RELATED STUDIES GROUP 2103 HISTORICAL STUDIES
Other comments	Loss of LIEF funding has threatened the viability of DAAO. They have pivoted successfully to an interoperable, user supported service, maintaining in a 'steady-state', being sustained by a very small staff and volunteer advisory group. Without ongoing support functions of DAAO and accessibility of their services will deteriorate, and this resource will struggle to maintain relevance to their research community.

Digital Observatory	
Description:	QUT's Digital Observatory facilitates access to significant large data collections drawing from social media and content sharing platforms. This includes the Australian Twitter Collection, Australian Twitter News Index, the Digital Media Observatory and the Australian Music Observatory. Central to the Digital Observatory is the Australian Twitter collection, a collection of tweets from all identified Australian accounts, collected since 2006. The dataset comprises more than 2.4 billion tweets with more than one million new tweets per day, and a comprehensive map of follower/ followee network structures and community clusters. The Australian Music Observatory provides longitudinal datasets on Australian music consumption patterns across a number of formats, while the Digital Media Observatory provides a comparative dataset on the availability of digital media content in Australia and the US. These collections form a phenomenal research resource that can power cutting edge research into Australian communication patterns, responses to global and local crises, political movements, language use, sentiment analysis, and can even detect the spread of disease ahead of other sources of evidence. They are an invaluable, near-real-time record of Australian cultural production and consumption, demographic trends and changing use of language and media.
Governance/ Management	The Digital Observatory is housed in and managed by the QUT Institute for Future Environments in partnership with the Digital Media Research Centre (DMRC).
Access and interoperability	The DMRC has developed a robust platform for managing and using data held in these collections, but access is limited by copyright, privacy and legal ownership considerations for this content. Access arrangements for social media content are subject to copyright laws, privacy information and the commercial restrictions of those platforms that publish the data. The Digital Observatory is committed to the principles of open access, and delivers appropriate access for researchers and research projects by arrangement. Interoperability with platforms including the proposed Linguistics Data Commons, and the Trove Researcher portal would considerably expand the access and utility of this big data collection within new and powerful environments.
Users, research communities, other stakeholders	DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE GROUP 2001 COMMUNICATION AND MEDIA STUDIES GROUP 2002 CULTURAL STUDIES GROUP 2003 LANGUAGE STUDIES GROUP 2004 LINGUISTICS GROUP 2099 OTHER LANGUAGE, COMMUNICATION AND CULTURE DIVISION 19 STUDIES IN CREATIVE ARTS AND WRITING

	GROUP 1901 ART THEORY AND CRITICISM GROUP 1902 FILM, TELEVISION AND DIGITAL MEDIA GROUP 1903 JOURNALISM AND PROFESSIONAL WRITING DIVISION 16 STUDIES IN HUMAN SOCIETY GROUP 1601 ANTHROPOLOGY GROUP 1602 CRIMINOLOGY GROUP 1603 DEMOGRAPHY GROUP 1604 HUMAN GEOGRAPHY GROUP 1605 POLICY AND ADMINISTRATION GROUP 1801 LAW
Other comments	Social media and internet research are emergent areas of humanities and social sciences. Australian leadership in this area is critical, to ensure cultural specificity and benefit to our community from this new area of research. There is great untapped potential in social media research, including observations on community wellbeing, the ways that messages are dispersed through social media, including in crisis response, or when there is community division, and better understanding of the practices of new media producers and new industries using these platforms.

Appendix 4: Consultation and relevant events:

Date	Meeting with	About
2020-03-17	s 47F	Consultation on HASS RDC Discussion Paper
2020-03-11		Consultation on HASS RDC Discussion Paper
2020-02-27		Consultation on HASS RDC Discussion Paper
2020-02-27		MAGDA and Data61
2020-01-29		Research and use of data in demography
2020-01-29		PARADISEC, progress on HASS RDC
2020-01-09		Improved research access to government data
2020-01-09		Indigenous data sovereignty, cultural safety, FAIR data
2019-12-20		Data linkage, sensitive data, PHRN and the humanities community
2019-12-18		Data linkage, demographic data proposal
2019-12-17		Digital Art History, Digital Humanities methods as taught at ANU
2019-12-06 /07	Knowledge Creation in the 21st Century: Approaches to Open, Digital Scholarship A Canadian-Australian Partnership for Open Scholarship (CAPOS) Gathering	Opportunities and challenges in open scholarship for the humanities; Canadian Australian collaboration
2019-12-05	s 47F	Integrated social sciences research infrastructure, Life Course Centre, ISSR
2019-12-03		RO Crate, OFCL, plans for HASS Research Data Commons

2019-12-03	s 47F	Archaeological data repositories and HASS research at Macquarie
2019-12-02		Progress on HASS RDC
2019-11-28		Integrated Research Infrastructure for Social Sciences
2019-11-27		Review of Nationally Significant Collections and Databases Scheme
2019-11-22		QUT and HASS research support
2019-11-21		AURIN
2019-11-20		GLAM Peak, GLAMR and cultural collections
2019-11-19		National and State Libraries
2019-11-19		APO
2019-11-18		AIATSIS collections and involvement in HASS RDC
2019-11-14		Media Archives Project, Centre for Media History
2019-11-05		HASS RDC
2019-10-31		Learning analytics and data sharing
2019-10-31		Time-Layered Cultural Map LIEF project

2019-10-30	s 47F	ACOLA involvement in HASS RDC consultation
2019-10-29		Creative Industries and Analysis and Policy Observatory
2019-10-22 /23	eResearch Australasia 2019: Digital Humanities and Indigenous Data stream; Social Sciences Stream	
2019-10-21	ARDC Data and Services Summit	
2019-10-18	s 47F	AustLII
2019-10-18		AustLit
2019-10-18		University of Melbourne Social and Cultural Informatics Platform, HASS research support
2019-10-17	Big Data in the Humanities: A Symposium	
2019-10-16	s 47F	Social Sciences
2019-10-15		Architecture collections and data
2019-10-15		CSIRO Knowledge Networks and HASS data collections
2019-10-12	Project team, Humanities Arts and Social Sciences Data Enhanced Virtual Laboratory	
2019-10-11	s 47F	Australian Data Archive
2019-10-10		AusStage
2019-10-10		Social History, Prosecution Project, Founders and Survivors

2019-10-09	s 47F	Architecture research and data
2019-10-08	Workshop: Social media, ethics, and data, University of Adelaide/ Australian and New Zealand Communication Association	
2019-10-04	s 47F	
2019-09-27		Tinker- HASS Data Enhanced Virtual Laboratory Project
2019-09-10		Trove and the HASS Research Data Commons
2019-09-10		Discussion of shared interests and planning for consultation at the Academy of Humanities Symposium
2019-09-09	Languages Data Commons of Australia Working Group	Progress plans for a Data Commons for linguistics.
2019-09-03	Emerging and Transformative Technologies in Art, Architecture and Design, UniSA	Introduction to new digital facilities in the Art, Architecture and Design School, and discussion of collaborations enabled by those tools. .

Group consultations:

- 15 November 2019, State Library of Queensland: Researcher consultation held in conjunction with the Australian Academy of Humanities annual symposium (click for consultation notes).
- 28 February 2020, Department of Education, Skills and Employment, Canberra: Consultation held by ARDC/ACOLA (click for consultation notes).
- 19 March 2020, Virtual consultation
- 24 March 2020, Virtual consultation

References

Alexander, Jane, 2019 "The First Anniversary of CMA Open Access: Benefiting People Now and Forever", Medium. https://medium.com/cma-thinker/the-first-anniversary-of-cma-open-access-benefiting-people-now-and-forever-9f3b70893534 (accessed 28 January 2020)

All European Academies, 2020. *ALLEA Report: Sustainable and Fair Data Sharing in the Humanities*, https://allea.org/portfolio-item/sustainable-and-fair-data-sharing-in-the-humanities/ (accessed 25 February 2020)

Alluvium, 2016, *Assessment of the Atlas of Living Australia's impact and value*. Report produced for the CSIRO, 50pp.

Arts Council England 2015 The Designation Scheme, Guidance for Applicants https://www.artscouncil.org.uk/supporting-collections-and-archives/designation-scheme (accessed 21 November 2019)

Arts Council England 2016 Pearls and Wisdom, Arts Council England's vision for the Designation Scheme for Collections of National Significance https://www.artscouncil.org.uk/supporting-collections-and-archives/designation-scheme (accessed 21 November 2019)

Asmi, Ari, Ryan, Lorna, Salmon, Emmanuel et al. 2019. International Research Infrastructure Landscape 2019 (Version 1). Zenodo. http://doi.org/10.5281/zenodo.3539254 (accessed 3 March 2020)

Australian Government, 2016 *2016 National Research Infrastructure Roadmap*, Commonwealth of Australia, Canberra.

Australian Government, 2016. *Facilities for the Future, Underpinning Australia's Research and Innovation: Government Response to the 2016 National Research Infrastructure Roadmap Research Infrastructure Investment Plan*, Commonwealth of Australia, Canberra.

Carroll, SR, et al. 2019. "Indigenous Data Governance: Strategies from United States Native Nations". *Data Science Journal*, 18: 31, pp. 1–15. DOI: https://doi.org/10.5334/dsj-2019-031

Commonwealth of Australia, Department of the Prime Minister and Cabinet, 2019, Data Sharing and Release Legislative Reforms Discussion Paper, https://www.datacommissioner.gov.au/resources/discussion-paper (accessed 21 Jan 2020)

Commonwealth of Australia, Department of the Prime Minister and Cabinet, 2018, The Australian Government's response to the Productivity Commission Data Availability and Use Inquiry https://dataavailability.pmc.gov.au/sites/default/files/govt-response-pc-dau-inquiry.pdf (accessed 21 Jan 2020)

Emery J, Boyle D, 2017 'Data Linkage' in *Australian Family Physician*, Volume 46, No.8, 2017 Pages 615-619

Harron K, Dibben, C et al. 2018. 'Challenges in administrative data linkage for research' in *Big data and society*, December 2017, vol. 4, issue 2 https://doi.org/10.1177/2053951717745678 (accessed 23 March 2020)

Humanities, Arts and Social Sciences Data Enhanced Virtual Laboratory, 2019 *Call for a National Digital HASS Research Framework: Strategic Priorities to effectively meet the digital needs of Australian HASS researchers* https://tinker.edu.au/call-for-a-national-digital-hass-research-framework/ (accessed 22 January 2020)

Library of Congress, 2019 "Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data: Funding to Allow LC Labs to Pilot Infrastructure for Digital Research at Scale", https://www.loc.gov/item/prn-19-098/library-receives-1m-mellon-grant-to-experiment-with-digital-collections-as-big-data/2019-10-04/?loclr=twndi (accessed 28 Jan 2020)

Mihelcic, Joanne, 2019 "Digital Humanities Case Studies: Lessons learned and Recommendations", a report produced for the Humanities Arts and Social Sciences - Data Enhanced Virtual Laboratory (available on request)

National Health and Medical Research Council, Australian Research Council and Universities Australia, 2018 *Australian Code for the Responsible Conduct of Research 2018*. Commonwealth of Australia, Canberra

National Health and Medical Research Council, Australian Research Council and Universities Australia, 2019. *Management of Data and Information in Research: A guide supporting the Australian Code for the Responsible Conduct of Research*. Commonwealth of Australia, Canberra

Productivity Commission 2017, Data Availability and Use, Report No. 82, Canberra

Russell, Roslyn & Winkworth, Kylie 2009 Significance 2.0: A guide to assessing the significance of collections 2nd edition. Collections Council of Australia

Social Sciences and Humanities Open Cloud, 2019, "D3.1: Report on SSHOC (meta)data interoperability problems" https://www.sshopencloud.eu/d31-sshoc-report-sshoc-data-interoperability-problems (accessed 22 January 2020)

The Official Statistics System, 2007. Principles and Protocols for Producers of Tier 1 Statistics 2007. http://archive.stats.govt.nz/about_us/who-we-are/home-statisphere/tier-1/principles-protocols.aspx (accessed 12 January 2020)

Whyte, A. & Wilson, A. 2016. "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides (accessed 12 January 2020)