Australian Research Data Commons

HASS Research Data Commons and Indigenous Research Capability project plan

IRISS – Integrated Research Infrastructure for Social Science

Revision History

Version	Date	Editor	Summary of changes
1.0	6 September 2021	Steven McEachern	First release
1.1	7 September 2021	Steven McEachern	Addition of draft sections 11-15
2.0	29 September 2021	Steven McEachern	Revision following public consultation

Australian Research Data Commons

1. Project title

IRISS – Integrated Research Infrastructure for the Social Sciences

2. Lead contact

The organisation with whom ARDC will contract.

- First name: Steven
- Last name: McEachern
- Organisation: Australian National University
- Group/Department: Australian Data Archive, ANU Centre for Social Research and Methods
- Job title: Director, ADA
- State: Australian Capital Territory
- Email: steven.mceachern@anu.edu.au
- Phone number: 0261252200
- ORCiD (optional): https://orcid.org/0000-0001-7848-4912

3. Proposal summary

Provide a summary of the aims, significance (or context), and expected outputs of the Project.

Write your Proposal Summary simply, clearly and in plain English. Avoid the use of acronyms and specialised terminology associated with a particular domain.

The Proposal Summary may be published to the general public as part of ARDC announcements.

Australia's social science research infrastructure needs to be better integrated and interoperable in order to keep up with leading edge global research and research infrastructures currently operating in the US and Europe. Much of the existing Australian infrastructure is underdeveloped, too small, has limited integration with other research infrastructure and is not well coordinated. Key existing social science data infrastructure have limited resources currently available to integrate or expand their significant existing data resources, and user base. A significant proportion of the available data also needs to be kept confidential and has serious privacy considerations. These constraints have led to a relatively fragmented infrastructure landscape for Australian empirical social science researchers.

The Integrated Research Infrastructure for the Social Sciences (IRISS) is intended to address this fragmentation, establishing a new foundation for integration of data, analysis and platforms for social science research in Australia. The starting point for this infrastructure is a core foundation of data - it's acquisition, documentation, harmonisation and dissemination for re-use. ANU have for 40 years maintained critical data infrastructure, the Australian Data Archive, to meet the needs of social science research. This project will extend and enhance that critical data infrastructure, by

developing integrated data and research infrastructure to meet the needs of SS researchers primarily and related disciplines that build their evidence base on longitudinal, demographic and geo-spatial data.

The core elements of this integrated infrastructure include:

- a foundational infrastructure for the acquisition, storage, documentation and dissemination of social science data
- extensible systems for the capture, documentation, preservation and analysis of data in near to real-time; and
- effective management of metadata and
- research environments for spatial and temporal data access, analysis and visualisation
- Capabilities for data linkage and integration across different data providers, using multiple data sources

The activities to be undertaken in this project, as part of the HASS-I program, will establish the foundations of the IRISS infrastructure, focussing on support for quantitative social science in Phase One (2021-23). IRISS will focus on the following project objectives, through a series of six coordinated work packages:

- Establishing a **coordinated** governance and integration model for the provision of data and infrastructure in the social sciences and related disciplines in Australia
- **Enhancing** the research capacity of Australian social science researchers, through the development of tools and services to enable the creation, dissemination and use of quantitative and qualitative social science data sources
- **Enabling** a cost-effective and accessible data integration environment (for lower risk data)

4. Project partners

Provide a short profile of the project partners that are collaborating on the proposal (this may be used in comms generated by ARDC).

Each of the partners in the IRISS project bring core resources to the project in one or more areas including research capacity, data, tools and services and training. The partners also service a broad user base within the social science community in Australia, with 7500 ADA users and 15,000 AURIN users from across the university, government and private sectors in Australia and around the world.

Australian Data Archive, Australian National University

The Australian Data Archive (ADA) provides a national service for the collection, preservation and dissemination of research data for secondary analysis by academic researchers and other users. The archive is based in the ANU Centre for Social Research and Methods (CSRM) at the Australian National University (ANU). ADA was established at the ANU in 1981 with a brief to provide a national service for the collection and preservation of digital data relating to social, political and economic affairs and to make these data available for further analysis.

ADA provides the only comprehensive social science data collection in Australia, with a catalogue of over 5000 data files across 1600 data sets, with a user base of over 7500 users from Australian and international universities, public sector organisations and the broader community. ADA hosts data from Australian surveys, opinion polls and censuses and includes data from other countries within the Asia Pacific region. ADA provides specialist services for subject areas across the social sciences, including politics, criminology and economics, and supports quantitative, qualitative, time series and panel data, and historical statistics.

The Australian National University (ANU) Centre for Social Research and Methods (CSRM) was established by in November 2014 to provide national leadership in the study of Australian society. CSRM has a strategic focus on: the development of social research methods, analysis of social issues and policy, training in social science methods, and providing access to social scientific data

Institute for Social Science Research, The University of Queensland

The Institute for Social Science Research at the University of Queensland is one of the largest social science research institutes in Australia, with more than 100 leading researchers, policy experts, and management professionals whose multidisciplinary approach delivers a broader perspective to the challenges our clients face. ISSR was established in 2007 to showcase UQ's strengths in social science research, to coordinate and integrate related research activities, and to position the University as a national leader in applied multidisciplinary social science.

ISSR is an international leader in advanced interdisciplinary and evidence-based social science research, and works collaboratively with government and the private and notfor-profit sectors on pressing social science challenges across four key impact areas: Social Policy and Practice; Health; Education; and Innovation and Technology.
This research is underpinned by cutting-edge social science methodologies including advanced data analytics, participatory and innovative qualitative research, observational and biometric

measurement techniques, experimental research designs, and the design and implementation of social intervention evaluations. Our strong focus on codesigned research, and commitment to ongoing training and development, provides significant engagement opportunities for our staff, postdoctoral students and industry partners.

ISSR is the administrative headquarters for the Australian Research Council Centre of Excellence for Children and Families over the Life Course, an international collaboration of 22 organisations working to identify the drivers of deep and persistent disadvantage and develop innovative solutions to address it. ISSR also hosts a node of the Australian Research Council Centre of Excellence for the Digital Child, as well as a site of the Centre for Social Data Analytics, based at Auckland University of Technology

Australian Urban Research Infrastructure Network (AURIN)

The Australian Urban Research Infrastructure Network (AURIN) is a National Collaborative Research Infrastructure Strategy (NCRIS) capability providing eInfrastructure and expert eResearch support for urban, regional and social science researchers in academia, government and industry. AURIN facilitate the development, deployment and long-term support of advanced data, analytical methods, simulation models and visualisation capability for the adoption of high-impact research within government and industry across Australia. These analytical tools and support services streamline access to health, transport, housing, economic, land use, demographic, and an extensive catalogue of other data — all integrated to work together.

AURIN works with more than 100 institutions including: Australia's leading universities, researchers, data custodians and government agencies, and has a user base of over 15,000 users. More than 500 people from the research, policy and practice communities are engaged in the AURIN project and more than 100 people are involved in AURIN committees and expert groups. The AURIN collaborative network simplifies interactions between researchers, data providers, policy makers and practitioners and services the growing demand for open-source access and evidence-based decision-making.

The Melbourne Institute, The University of Melbourne

The Melbourne Institute is Australia's pre-eminent economic and social policy research institution. It undertakes high quality, independent and impartial applied research, and contributes to the development of public policy in Australia. Established in 1962, the Institute provides high calibre, peer-reviewed levels of research to government, business and community groups. It is internationally renowned for its HILDA and MABEL Surveys, as well as its measurement of economic and social indicators.

Members of the MI data and analytics team are highly skilled and come from a variety of backgrounds spanning industrial and academic research, in the fields of applied economics and social sciences, statistics, engineering, data science and machine learning. The MI data and analytics team members support project delivery and enable researchers to derive meaningful insights from multiple sources of data, ranging from survey and administrative data to nontraditional sources. The team also provide technical assistance in data management practice and data storage solutions.

Australian Consortium for Social and Political Research, Incorporated (ACSPRI) (to be confirmed)

The Australian Consortium for Social and Political Research Incorporated (ACSPRI) is a not-for-profit organisation, formed in 1976 with the broad aim of the promotion and enhancement of social science research and methods in Australia. ACSPRI's objectives are to:

- Facilitate access to Australian and overseas sources of computer-readable social science data;
- Encourage and support activities and procedures which enhance access to and use of social science data;
- Collect and disseminate information relating to social science data; and
- Encourage and support teaching and research in the social sciences.

ACSPRI provides regular training programs in social research methods and research technology, with annual programs in Melbourne, Sydney, Canberra and Brisbane. ACSPRI also runs a survey research centre, specialising in data collection for academic social science research, as well as methodological research and the development of software and tools. In addition, we work with members on other infrastructure development as required.

5. Project team roles and responsibilities

The following table defines the roles and responsibilities of key stakeholders and staff throughout the implementation of the project. Please add more rows as required to describe each staff member. Include any external parties that have a role within the project.

Name	Project role	Organisation	Responsibility
Steven McEachern (ANU)	Project Lead	Australian National University	Project leadership, strategic direction, leadership of work packages
Project Manager (to be appointed)	Project Manager	Australian National University	Overall project management and work package coordination
Matthew Gray	Lead institution representative	Australian National University	Steering Committee, leadership of Demonstrator Project 3
Mark Western	Partner Lead	University of Queensland	Steering Committee, leadership of work package 4 (GeoSocial) and Demonstrator Project 1
Stuart Barr	Partner Lead	Australian Urban Research Infrastructure Network	Steering Committee, technical leadership (GeoSocial)
A. Abigail Payne	Partner Lead	University of Melbourne	Steering Committee, leadership of Demonstrator Project 2 and work package 6 contributor
Adam Zammit	Stakeholder, expert advisor	Australian Consortium for Social and Political Research, Incorporated	Steering Committee, work package 5 contributor
Len Smith	Expert advisor	Indigenous Data Network	Steering Committee, technical advisor (Indigenous Data)
ABS representative (to be confirmed)	Expert advisor	Australian Bureau of Statistics	Steering Committee, technical advisor (Classifications and Official Statistics)
DSS representative (to be confirmed)	Expert advisor	Department of Social Services	Steering Committee, technical advisor (Public sector survey and administrative data)
Staff roles	Various	Various	Specific roles working on activities within IRISS work packages are specified in each work package below.


Australian Research Data Commons

6. Project objectives

List the objectives of this project. Include a description of any existing infrastructure, data collections, platforms, services and tools that are being leveraged as input for this project. (500 words)

Phase One of IRISS, to be established through the HASS-I activities in this project, has three core objectives:

- Establishing a **coordinated** governance and integration model for the provision of data and infrastructure in the social sciences and related disciplines in Australia
- **Enhancing** the research capacity of Australian social science researchers, through the development of tools and services to enable the creation, dissemination and use of quantitative and qualitative social science data sources
- **Enabling** a cost-effective and accessible data integration environment (for lower risk data)

The project is oriented around improved integration across the research lifecycle of the social sciences (Figure 1).
To achieve the project objectives, each work package in the IRISS project is aligned with

a) a phase of the research lifecycle, as illustrated in Figure One, and
b) one or more of the three IRISS objectives (detailed in the work package descriptions in Section 7)

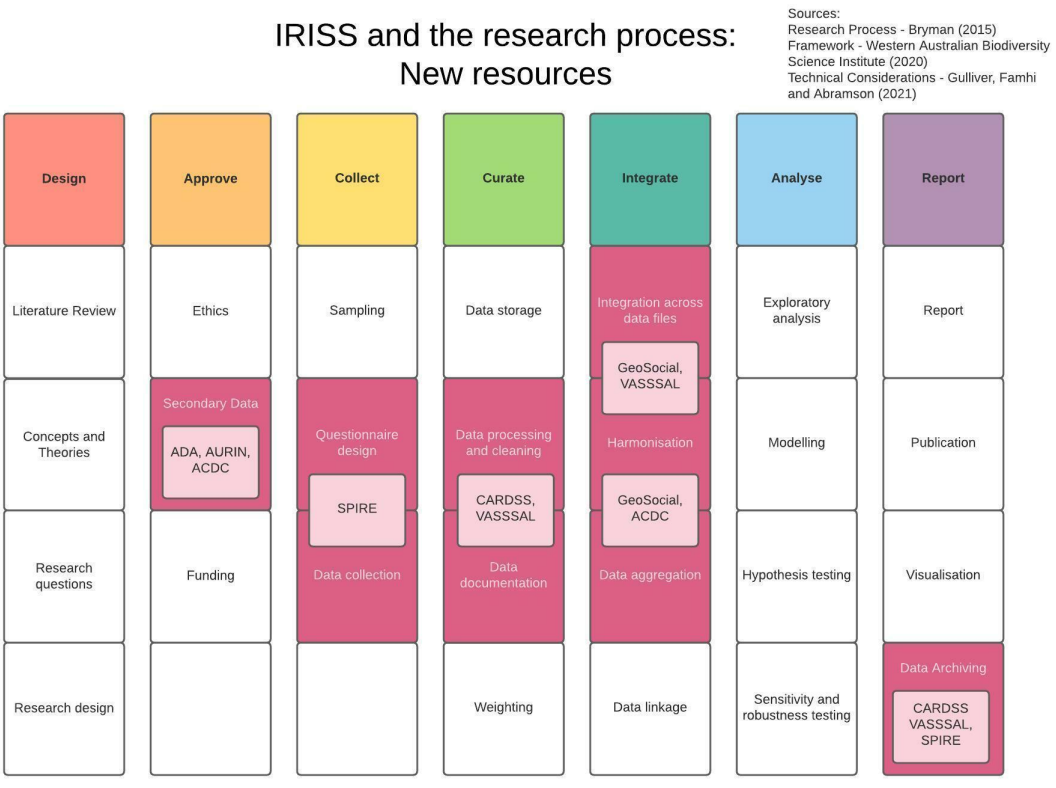

Figure 1 IRISS Research Process model and work packages

As an integration program, it is also critical that IRISS leverage key national research infrastructure. To this end, IRISS will incorporate data, tools and services from the following established services and facilities:

- Core social science data sources, including the Australian Data Archive, Australian Bureau of Statistics and Australian Urban Research Infrastructure Network
- Web-based data collection tools and services, include the ACSPRI LimeSurvey service and commercial services such as Qualtrics
- National and international data discovery, vocabulary and identifier services (Australian Research Data Commons, CESSDA - Consortium of European Social Science Data Archives) to enable data and metadata discovery across IRISS services;
- Data storage facilities (National Computational Infrastructure) providing high availability and large-volume data storage and back-up facilities;
- Secure data access infrastructure (CADRE) and secure high capacity networks (AARNet) for transfer of data between researchers and facilities, and to enable remote access to secure facilities;
- Training and education capacity in social science research methods (ANU, UQ, UniMelb, ACSPRI) and digital skills and data management (ARDC)

The expected outputs of the project include:

- A governance model and roadmap for future investments in Australian social science infrastructure
- Access to broader and higher quality data from academic institutions and the public sector
- Enhancement and integration of data collections of national significance (e.g. Australian Census, Australian Survey of Social Attitudes)
- Integration capabilities for multiple data sources - by spatial, temporal dimensions and units of analysis

7. Project activities

The IRISS Project Activity will consist of six work packages to be completed between December 2021 and June 2023. Five of these packages are aligned with the research process model introduced in the Project Objectives (Section 6, Figure 1) above. Package One will coordinate project management and technical and data implementation activities across the work package. Packages 2 and 3 establish core data integration services – VASSSAL and GeoSocial – for conceptual, spatial and temporal data integration support, while Package 4 will implement a series of demonstrator projects testing the services and illustrating the implementation of the project infrastructure in applied social science research settings. The final two work packages, SPIRE and CARDSS (Packages 5 &6) provide additional data collection and curation services for the improved management of research data throughout the research process.

Work package 1: IRISS Project Coordination

A. The objective of the work package

Timeframe: 18 months

Budget:

1.A: Project Management: $563,905

1.B: Technical Management: $440,762

Project objective: Coordinating

This work package will provide overall project coordination to provide alignment of work package activites within the project, and coordination with other HASS-I projects and external activities. This coordination will be in two areas:

1. Project management and coordination: strategic direction of the project, operational project management and external relationships and communications
2. Technical management and integretation: project-wide technical architecture design and implementation, IRISS front-end interface development and development of IRISS web presence

B. The relationship between project components

This package will provide overall project and technical coordination for all IRISS activities. As such, it will have established relationships with all project work packages.

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

Project management team:

- Project lead (ANU, 0.5FTE, 18 months)
- Project manager (ANU, 0.5FTE, 18 months)
- Project coordinator (ANU, 1.0FTE, 18 months)

Technical management team:

- Technical architect (ANU, 0.8FTE, 18 months)
- IRISS data architect (ANU, 0.8FTE, 18 months)
- Interface developer (ANU, 0.5FTE, 15 months)

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

Deliverables:

- Project work plan (January 2022)
- User engagement strategy and plan (December 2022)
- IRISS technical and data architectures (December 2022)
- IRISS technical and data integration strategy and plans (December 2022)
- Project progress report (December 2022)
- Project final report (June 2023)

Work package 2: VASSSAL – Vocabulary Access Service for Social Science in AustraLia

A. The objective of the work package

Timeframe: 18 months

Budget: $201,685

Research processes: Integrate, Report

Project objective: Coordinating

Classifications and vocabularies form a core part of social science research infrastructure in Australia. Major classification systems, particularly those provided by the Australian Bureau of Statistics (ABS), are regularly adopted and used in multiple data collections, and provide a core opportunity for research data discovery, analysis, reuse and integration.

However, although classifications and vocabularies are accessible to researchers, the publication formats are not readily usable by the current and emerging collection and analysis tools used by HASS-I researchers. ABS classifications exemplify these concerns – a researcher using such classifications in a research project will need to expend considerable resources to reformat the classification, to enable use both in their own data analysis projects, and in processes of integration between ABS data, ABS classifications, and other data. This resource intensive process of transformation for integration and analysis is repeated across every such research project. And with every such process there is the possibility that errors are introduced into the transformed data.

The VASSSAL work package is intended to address this concern, by establishing **a core service for the creation, dissemination and reuse of classifications and vocabularies** in Australian social science. Working with partners at the ARDC Research Vocabularies Australia (RVA) and the ABS, the VASSSAL package will establish a pilot vocabulary service hosted through RVA to support core social science classification and vocabulary requirements, such as statistical classifications, question libraries and survey response formats. The work team would then partner with the ABS to pilot activities to produce versions of nominated ABS data and classifications, in formats enabling access and reuse across the research sector. Emphasis would focus on activities which will provide infrastructure to boost research efficiency, place minimal demand on ABS resources, and drive increased reuse of ABS data and increased adoption of ABS classifications. (The ABS have been consulted on initial options for this project and have indicated their interest in involvement).

B. The relationship between project components

Inputs: VASSSAL will takes classifications and vocabularies from ACDC (Census Demonstrator project) and SPIRE, and integrate with international vocabulary services such as the CESSDA vocabulary service and DDI Alliance vocabularies

Outputs: VASSSAL vocabularies will be accessible by all other services, but will particularly provide outputs for SPIRE (survey questions and response domains) and GeoSocial.

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

Business analyst (ANU, 0.5FTE, 12 months)

Database developer (ANU, 0.5FTE, 15 months)

Data architect (*contribution as part of WP1B*)

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

- User Requirements – researcher input for demonstrators (Spatial, Sensitive and Census Data) (Q1)
- Technical Report – selection of two statistical classifications (for high impact and mass research reuse) (Q1)
- Service Design – Vocabulary Service – technical and data architectures and integration (ADA, AURIN & RVA) (Q1-2)
- Service Pilot – Vocabulary Service (two classifications and two demonstrators) (Q3-5)
- Technical Report – recommendations (based on user testing) and next steps for transition plan for move to operation and expansion (Q5-6)

Work package 3: GeoSocial data integration service

A. The objective of the work package

Timeframe: 18 months

Budget: $987,933

Research Process: Integrate

Project objective: Enabling

The objective of the GeoSocial Australia work package is to establish innovative spatially integrated approaches to social science research that can directly inform local and national efforts to improve people's lives. By linking the geospatial statistical data of the Australian Census to the person-centred primary data and administrative records in our nation's largest longitudinal surveys, we will empower a large cross-disciplinary social research community in Australia to identify patterns, make predictions, and inform social policy using rich integrated geosocial data. As the Academy of Social Sciences in Australia has argued, the capacity of Australian social research for research impact depends on the ability "increasingly to predict patterns and outcomes at large and small social and spatial scales" (2018).

In this work package, The University of Queensland's Institute for Social Science Research (ISSR), the Australian Urban Research Network (AURIN), and the Australian Data Archive (ADA) will design a search, retrieval and integration environment that can generate data products that integrate people, place, time and space. These multi-dimensional data reflect the breadth and depth of variables that sociologists, geographers, social statisticians, epidemiologists and others have been demanding.

To achieve this, we will collaborate with government data custodians and prospective end-users to integrate three significant national data holdings: 1) the ABS Census 1981-2016; 2) key longitudinal surveys in the ADA Dataverse Project; and 3) the Australian Census Longitudinal Dataset 2006-2016. Together, these data comprise variables with the capacity to reflect rich spatial and social trajectories for more than 115,000 individuals. We will develop an interface that facilitates effective interaction between the ABS census data and records in the ADA Dataverse in order to support secure integration of person-based and place-based data on demand. We will adapt recognised data management protocols to support strong data stewardship and to manage potential social licence and data privacy concerns, such as confidentiality, reidentification and data utility.

With this new infrastructure, the national capacity to ascertain important spatio-temporal changes for place-based planning about wide-ranging issues relating to population health, social wellbeing, and community cohesion will be enhanced. It will reduce duplication in labour-intensive efforts by individual researchers to source and combine data that can illuminate place-based population dynamics and it will expedite research results. GeoSocial Australia will achieve high quality research outcomes that demonstrate the feasibility and value of spatially integrated social science data, which can be extended to other data types and structures in future.

B. The relationship between project components

The GeoSocial service will use vocabularies and classifications provided through VASSSAL to support integration, and the Census data provided through the Australian Census Digital Collection as the source data for integration.

New data assets derived from the GeoSocial service will be analysed using in the Spatial Data Analysis project in the Demonstrators work package.

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

Work package lead (UQ, 0.8FTE, 18 months)

Data curator (UQ, 1.0FTE, 18 months)

Software engineer (AURIN, 1.0FTE, 18 months)

Database developer (AURIN, 1.0FTE, 18 months)

Technical architect (contribution as part of WP1B)

Interface developer (contribution as part of WP1B)

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

- User Requirements – researcher input for demonstrators (Spatial & Sensitive Data) (Q1)
- Technical Report – data architecture and protocols (based on three datasets and user requirements) (Q2)
- Service Design – Transformation Service – technical and data architectures and integration (ADA & AURIN) (Q1-2)
- Service Pilot – Transformation Service (test with three datasets and two demonstrators) (Q3-5)
- Technical Report – recommendations (based on user testing) and next steps for transition plan for move to operation and expansion (Q5-6)

Work package 4: IRISS Demonstrator projects

A. The objective of the work package

Timeframe: 12 months

Budget: $277,124

Research processes: All

Project objective: Coordinating, Enhancing, Enabling

A key objective of the IRISS project is the enhancement of the capacity of Australian social science researchers to creation, dissemination, integrate and use quantitative and qualitative social science data sources to generate new insights.

To this end, the Demonstrator work package will conduct three small scale demonstration projects to establish and illustrate the integrative capabilities of the combined services in IRISS. These two projects are intended to establish the fitness for purpose of the services developed, as well as to establish requirements for Phase Two of IRISS – supporting varied forms of data analysis (such as spatial data analysis, use of sensitive data in suitable trusted facilities), and enabling integration of analysis tools and environments.

The three demonstrator projects include:

1. The **Spatial data analysis demonstrator** will be lead by the University of Queensland. This demonstrator will leverage the census data outputs of ACDC, and the integration services of the GeoSocial service, to establish a new data product, and apply person-based and place-based methods to the product.
2. The **Sensitive data analysis demonstrator** will be lead by the University of Melbourne. This demonstrator will assess the new outputs of IRISS services (such as new surveys generated through SPIRE, and curation tools from CARDSS) to assess their suitability for use in sensitive data environments, which are subject to higher levels of data and information security, ethics assessment and confidentiality risks. The project will leverage the Melbourne Institute Data Laboratory to conduct these analyses, providing a pilot study for future sensitive data support within IRISS.
3. The **Australian Census Digital Collection demonstrator** will be lead by the Australian National University. The use of census data forms the foundation of much major data-oriented research in Australia, and represents the key reference data asset in Australia on human population and settlements. The availability of this national asset is however less than is needed for such a significant asset. A significant proportion of population data is located in archives and libraries, or digitised but embedded in formats largely inaccessible to researchers or to the Australian public. This work package will implement the first stages of the collection development program, to establish the Australian Census Digital Collection (ACDC), extending work ADA undertook in a national audit of the current state of small area census data in Australia supported by the ARDC in 2019. This work package will include three pilot activities to pilot development: (a) collection migration and development program to move content into fully machine actionable formats, (b) documentation preservation program to capture, preserve and disseminate the documentation associated with the census in each year of collection, and (c) a metadata creation program for the capture of metadata from the censuses to enable data integration and harmonisation, along with metadata reuse in other services

B. The relationship between project components

- Inputs: The demonstrator projects will utilise data, tools and services established across the other IRISS work packages.

- Outputs: The projects will generate new data and research outputs, which will be preserved and accessible through publication and data repositories (such as ADA and AURIN)

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

- Analyst, Demonstrator One (UQ, 0.5FTE, 12 months)
- Analyst, Demonstrator Two (UniMelb, 0.5FTE, 12 months)
- Analyst, Demonstrator Three (ANU, 0.5FTE, 12 months)

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

- User Requirements – researcher input for demonstrators (Spatial, Sensitive and Census Data) (Q1)
- Technical Report – overall data and technical architecture for integration across research lifecycle (based on user requirements) linking activity undertaken IRISS work packages, contingent requirements in the wider HASS RDC and ADA ANZLEAD & CADRE projects (Q2)
- Infrastructure Strategy and Plan – overall vision for IRISS as an integrating national research infrastructure for Social Science and allied disciplines (Q2)
- Service Design – Integration Service – data and technical architectures and integration (Q3-4)
- Technical Report – recommendations (based on user testing) and next steps for transition plan for move to operation and expansion (Q4)

Work package 5: SPIRE - Survey Project Integrated Research Environment

A. The objective of the work package

Timeframe: 12 months

Budget: $163,444

Research process: Collect, Curate

Project objective: Coordinating, Enhancing

Survey data collection and analysis one of the work-horse process for social science researchers. The advent of the web moved much of this data collection process into highly structured, web-based tools, and there are now major commercial services and open source tools providing secure, web-based survey data collection. These services often provide API-based access to both data and metadata, and there is a small ecosystem of open-source libraries developing to support interaction with specific services.

While this ecosystem is emerging, there are clear gaps in the ecosystem:

- Limited use of the APIs for survey creation
- Data generated by these services is largely still held within desktop and personal data holdings after export
- There is limited use of standardised libraries and vocabularies for either survey instrument creation, or of reuse of metadata (in the form of variable and question specifications) in future data collection projects

The SPIRE work package therefore aims to address these three gaps to establish an end-to-end processing and metadata support environment for survey data collection and archiving. The aims of the SPIRE system will be three-fold:

- Align survey data collection with vocabulary and question bank services (leveraging existing content from the VASSSAL vocabulary service)
- Harmonise archived content with available machine readable classifications and vocabularies (providing new content for VASSSAL from data archived with ADA and other facilities)
- Throughput of newly generated survey data into data processing environments (such as Cloudstor) and data archives and repositories for long term preservation and access

These three data and metadata flows are laid out in Figure 2.

B. The relationship between project components

The SPIRE facility will leverage VASSSAL and the concurrent ANZLEAD project, to connect existing survey question and variable components (such as question text and response domains) to survey data collection tools.

SPIRE will provide access to data processing tools established in the CARDSS curation service to support researcher data processing of survey data.

Data outputs of the SPIRE facility will be provided into existing data repositories such as the Australian Data Archive.

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

Software developer - R/Python (ANU, 0.5FTE, 12 months)

Data scientist (ANU, 0.5FTE, 12 months)

Project coordinator (contribution as part of WP1A)

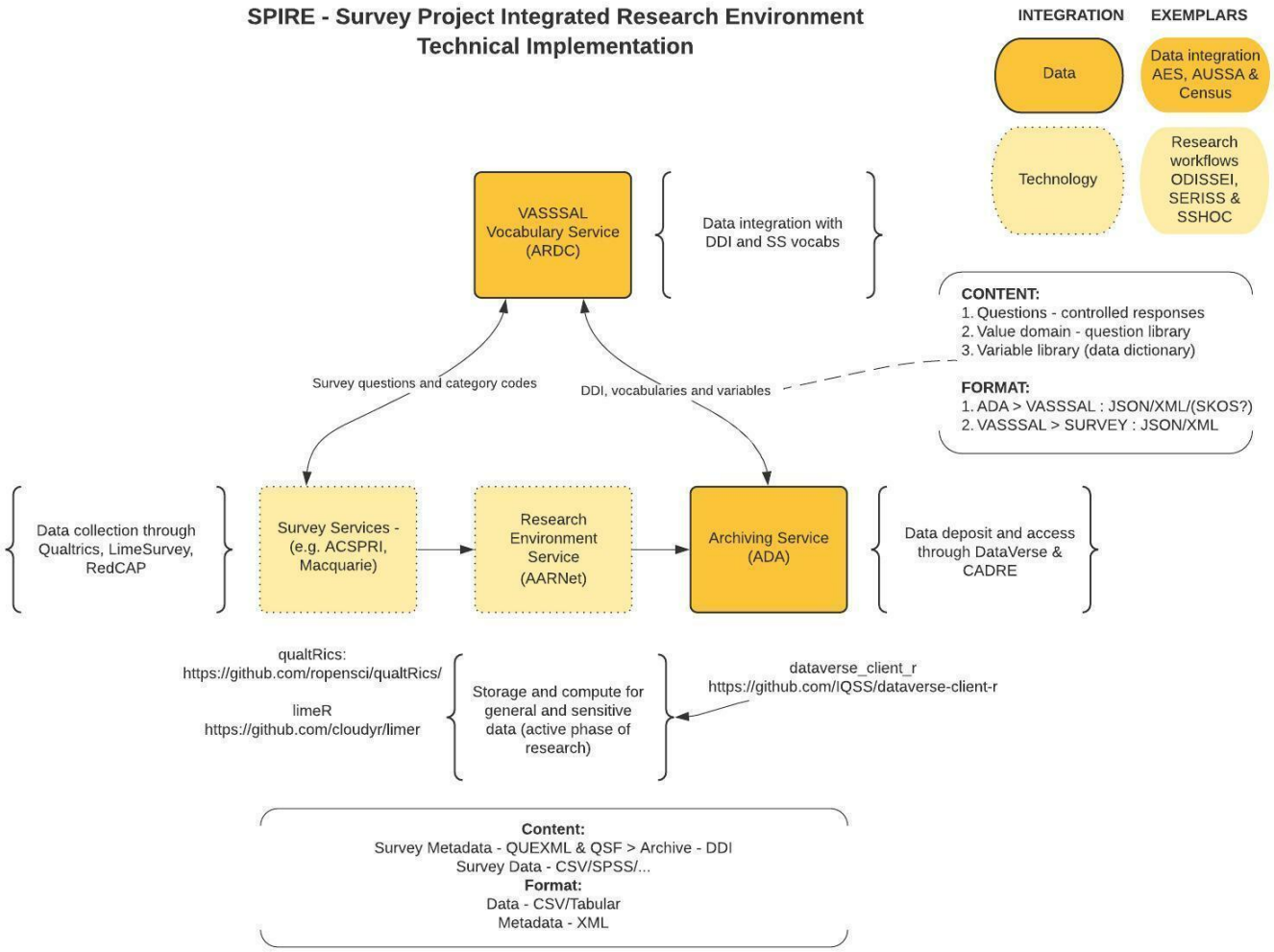

Figure 2 SPIRE technical implementation

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

- User Requirements – researcher input for demonstrators (Spatial & Sensitive Data) (Q1)
- Technical Report – data architecture and protocols (based on user requirements for demonstrators) (Q2)
- Service Design – API Service – technical and data architectures and integration (ADA, RVA, 1 Survey System) (Q3)
- Service Pilot – API Service – test with 1 Survey System (Q3)
- Technical Report – recommendations (based on user testing) and next steps for transition plan for move to operation and expansion (Q4)

Work package 6: CARDSS – Curation of Australian Research Data in the Social Sciences

A. The objective of the work package

Timeframe: 12 months

Budget: $219,124

Research processes: Curate, Report

Project objective: Enhancing

There is a strong tradition of reuse of secondary data sharing in the social sciences. Tabular data is highly amenable to sharing, with well-established data repositories such as the ADA, and major collections of public data such as the Australian Bureau of Statistics. The procedures for curating this data however have traditionally been manually organised, with the diversity of content areas making standardised processing difficult.

Increasingly however, the development of electronic data collection tools (see the SPIRE package above) makes the standardisation of processing more tractable. The availability of standard output formats for both data (CSV, Stata, SPSS) and metadata (DDI, XML) mean that there is more homogeneity in the content being created. In turn, this has the potential to enable more streamlined curation of quantitative tabular data, through the development of curation tools for the management of newly created data. Such tools could assess data and metadata quality (such as frequency distributions, variable and value labels), privacy and confidentiality risks, and standard curation and data management practices (such as file formats and naming conventions).

This work package will establish a standardised program library and training packages for the management of social science research data. The program library will be built using standardised processing tools (starting with the open source R program, and extending to other commonly used software such as Stata and SPSS), and will be made available to researchers through a shared statistical software libraries (e.g. R packages) for use in day to day data management activities. An associated set of libraries will be developed for the archiving and storage of data and metadata with repositories such as ADA and other HASS-I projects where relevant. The work package team will then develop training and communications materials to support the adoption and use of these libraries, delivered through online learning systems.

These libraries will also integrate with other IRISS work packages. Processing libraries will support the use of standardised vocabularies and classifications (such as national and international classifications from the Australian Bureau of Statistics) in the creation of metadata and variables, and the generation of new metadata for inclusion in the VASSSAL vocabulary server.

B. The relationship between project components

Inputs: Data and metadata generated by the SPIRE facility will be processed using the CARDSS curation tools. CARDSS will also enable reuse of standard classifications available in the VASSSAL service.

Outputs: The CARDSS program library will enable data archiving and storage in HASS-I repositories, including ADA. The library will also be used in curation activities occurring in other IRISS work packages, including SPIRE survey data, ACDC census data and the IRISS Demonstrator projects.

C. The estimated effort involved in delivery (role and FTE of allocated individuals),

Software developer - R/Python (ANU, 0.5FTE, 12 months)

Data scientist (ANU, 0.5FTE, 12 months)

Project coordinator (contribution as part of WP1A)

D. Who/what will determine the work package as complete and fit for purpose (deliverables).

- User Requirements – researcher input for demonstrators (Spatial & Sensitive Data) (Q1)
- Technical Report – data and technical architecture (based on user requirements two demonstrators) (Q2)
- Service Design – Curation Service – data and technical architectures and integration (ADA & RVA) (Q3)
- Service Pilot – API Service – test with one software output (Stata or SPSS) (ADA & RVA) (Q3)
- Technical Report – recommendations (based on user testing) and next steps for transition plan for move to operation and expansion (Q4)

8. Integration Component

Describe, in detail, elements of your proposal that will utilise RDC wide developments to be supported by the ARDC Integration Component.

1. Access, Authentication and Governance

- Enable research groups and communities that do not currently have AAF credentials to access HASS-I infrastructure including ADA and LDaCA.
- Implement AAF authentication as the preferred access model.
- This includes governance around the FAIR and CARE principles.

2. Consultation Phase

- Shared Model for engagement with communities.
- Develop evidence-based user requirements for modelling common services.

3. Compute: HPC and GPUs

- To be able to support the use of machine learning/AI in text analytics.

4. Skills and Training

- Project specific training
- Scope of what ARDC training exists

5. HASS RDC Roadmap

There is a need to evaluate and extend the outcomes of current HASS-I project activities to establish directions for further work as part of the National Research Infrastructure Roadmap. A shared work package would include the need to develop road maps for both components.

a. HASS RDC Roadmap
    i. Governance framework
    ii. Legal and cultural data access framework
    iii. Technical architecture
    iv. Building communities
b. Indigenous Research Roadmap

i. The application of Indigenous governance frameworks

ii. Legal/legislative requirements around data

iii. Technical architecture and the ongoing development of catalogue

iv. Indigenous Research Capabilities

6. Legal requirements for data management and infrastructure

This work package would begin the conversation across government around the legal requirements of data access and use.

9. Outcome and Impact

Provide use cases which demonstrate how the deliverables will support research programs.

Project use cases are suggested here aligned with the three demonstrator projects identified in Work Package 4. These use cases illustrate three core challenges for social science researchers in the use of data and services, that the IRISS program seeks to address.

Use Case 1: Spatial Data Analysis

A core element of many research and policy questions in the social sciences relates to the impact of a person's social context on their life experiences and outcomes. One important context in many domains is the impact of place – where someone lives, works and studies. Researchers across social science domains as diverse as sociology, geography, economics and public health regularly study the impact of place on life outcomes, using place-based measures such as population (e.g. density, diversity), environment (air quality, water quality), urban infrastructure (green space, housing) and public safety (crime, policing).

Studying such measures however requires extensive data integration of information about individuals with information about places – linking data across datasets with diverse content, formats, scales and geographies. Such integration is time-consuming, inconsistent and error-prone, requiring knowledge of both person and place-based content in order to study such problems. The use of spatial data also requires the application of analytic tools and methods suited to geographically clustered information such as spatial autocorrelation models (e.g. Logan et al., 2010). Streamlined services for bringing together person and place data into integrated data products will allow for significant improvement in the creation of integrated spatial data sets, and the analysis of place-based research questions, using techniques optimised for such data formats.

Logan, J. R., Zhang, W., & Xu, H. (2010). Applying spatial thinking in social science research. *GeoJournal*, 75(10), 15–27. https://doi.org/10.1007/s10708-010-9343-0

Use Case 2: Sensitive Data Analysis

The focus of social science research on humans and their lived experiences inherently involves the collection and analysis of sensitive data. While such data is either created for the purpose of research, or approved for use in research by custodians, such use is often subject to conditions of use imposed by data owners, participants and those acting in their stead such as human research ethics committees. These conditions then create challenges for researchers in conducting their research, as the transition of data from one stage of research to another is limited by physical and system security constraints.

Secure research environments, such as the Melbourne Institute Data Lab (MIDL) and the Australian Bureau of Statistics ABS Lab have emerged to address this challenge. The transition of data in and out of these environments, and the integration of sensitive data from multiple sources, continues however to be a challenge in studying complex, sensitive social science problems (e.g. Christen et al., 2020). Closer alignment across the Five Safes of data access (Desai et al., 2016) is therefore required. Leveraging the data integration tools established in the IRISS project

(GeoSocial and SPIRE), along with existing access methods (CADRE) and secure analysis facilities (MIDL), IRISS will demonstrate how closer system and data integration can accelerate research outcomes for users of sensitive data.

Peter Christen, Thilina Ranbaduge, and Rainer Schnell (2020). *Linking Sensitive Data Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer.

Use case 3: Census and historical tabular data

Australia's Commonwealth census data collection is now over 105 years old, with the completion of the 2016 Census. Censuses and musters have also been undertaken in the Australian colonies since 1838. A significant proportion of population data in Australia is human-readable, but located in archives and libraries, or digitised but embedded in formats largely inaccessible to researchers or to the Australian public. Even more challenging is the lack of ability to bring the data into a format suitable for researcher analysis and use. While data from 1996 onwards is available from the Australian Bureau of Statistics, data prior to this date is difficult to locate and access.

ADA has worked to digitise census content provided by the Australian Bureau of Statistics, but the current format and nature of this content has limitations for researchers. Data from 1911-1961 is embedded in PDFs and print materials. For the census data from 1966 to 1981, there are no corresponding maps or geocode files of the geographic coverage of these censuses, limiting the usefulness of the data for informing spatially-enabled research or policy.

Transformation of content from these collections into machine-actionable formats would enable the study of long-term transitions in the Australian economy, society and polity. Applications include economic history (e.g. Grosjean and Khatter, 2017), population mental health (e.g. Hanigan et al., 2012) and indigenous studies (e.g. Markham and Biddle, 2016). The Census data demonstrator project will evaluate text analytics and machine learning tools for the extraction of historical tabular data, to support current historical and demographic research, and establish future directions for the use of digitised materials in studying long-term social change.

Grosjean, P., & Brooks, R. C. (2017). Persistent effect of sex ratios on relationship quality and life satisfaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1729), 20160315. doi:10.1098/rstb.2016.0315

Hanigan, I., Butler, C., Kokic, P. & Hutchinson, M. (2012). Suicide and drought in New South Wales, Australia, 1970–2007. *Proceedings of the National Academy of Sciences,* Aug 2012, 109 (35) 13950-13955; DOI: 10.1073/pnas.1112965109

Markham, F. & Biddle, N. (2016). *Indigenous residential segregation in towns and cities, 1976–2016*. Census Paper No. 4, Centre for Aboriginal Economic Policy Research, Australian National University, Canberra.

NB. ARDC together with the partners will implement a process for reporting on usage and outcomes of the infrastructure we build together. Reporting on impact allows us to demonstrate the value of the programs, activities and services to the commonwealth, states, partners, research community, and broader society. Partners are required to support the ARDC in implementing an impact reporting framework .


Australian Research Data Commons

10. Governance

State who is accountable for assessing project performance, and what processes will they apply. Describe the steering committee that will oversee the project, its frequency of meetings, and list its proposed members. The ARDC HASS RDC Program Manager should be included in the governance body.

The IRISS Project Steering Committee will be established to oversee the implementation and performance of the project. The steering committee will be drawn from members of the project, the Australian social science research community, government agencies and non-government organisations (as consumers of social science research), and the ARDC as project partners.

The committee will be chaired by an independent chair, and provide approval of the key deliverables of the project – primarily the overall project deliverables outlined in work package 1, including project plan, periodic reporting and financial reports.

The proposed membership of the committee is as follows:

- Project lead (ANU)
- HASS RDC Program Manager (ARDC)
- Partner leads (ANU, UQ, AURIN, UniMelb and ACSPRI)
- Research community representatives (2)
- Government agency representatives (1 Commonwealth, 1 State)
- Non-government organisation representative (1)
- Academy of Social Science in Australia representative (1)
- Indigenous Data Network representative (1)

The independent chair of the Committee will be drawn from the members of the committee. Representatives employed by project partner organisations will not be eligible to act as chair of the Committee.

The committee will meet quarterly, with secretariat support provided by the project management group as part of Work Package 1. The project lead and chair of the IRISS steering committee will also participate in joint HASS-I program governance as required.