Australian Research Data Commons

‘People’ Research Data Commons

BRIEFING PAPER FOR CONSULTATIONS WITH NATIONAL STAKEHOLDERS

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

CONTENTS

EXECUTIVE SUMMARY

The ARDC is developing a suite of Thematic Research Data Commons[1] (Thematic RDCs) that scale up digital research infrastructure to meet Australia's future research needs, with two initial pilot Thematic RDCs being established in the 2022-23 financial year. The People Research Data Commons or People RDC is the first of the pilot Thematic RDCs and it will focus on health and biomedical research.

Health research is a national priority area, with the National Research Infrastructure Roadmap including 'Medical Products' as a priority area, and significant research funding across academia and industry applied to medical and health sciences. The People RDC is a vehicle for the ARDC and our national partners to collaboratively develop and deliver sustainable digital research infrastructure on a national scale in a strategic and comprehensive manner.

The ARDC is taking a conceptual model of a Thematic RDC and extending it to detailed models for the People RDC through stakeholder consultations.

This paper provides an overview of the People RDC. It outlines the conceptual model and proposes key data challenges that could be addressed by the People RDC to meet the needs of the national health research community. It sets the context for the consultation process that will examine the data challenges and the priorities to enable the development of an implementation plan for the People RDC.

Note: In the context of the pilot Thematic RDCs, People RDC is a working title for the purpose of this paper and the consultations.

[1] *https://ardc.edu.au/news/designed-for-the-future-thematic-research-data-commons/*

BACKGROUND

A Thematic RDC brings together closely aligned research domains, industry and government to establish and maintain fit-for-purpose data assets, best practices and digital research infrastructure in order to drive increased excellence, resilience, optimisation and agility in research and research translation for societal benefits and commercial opportunities.

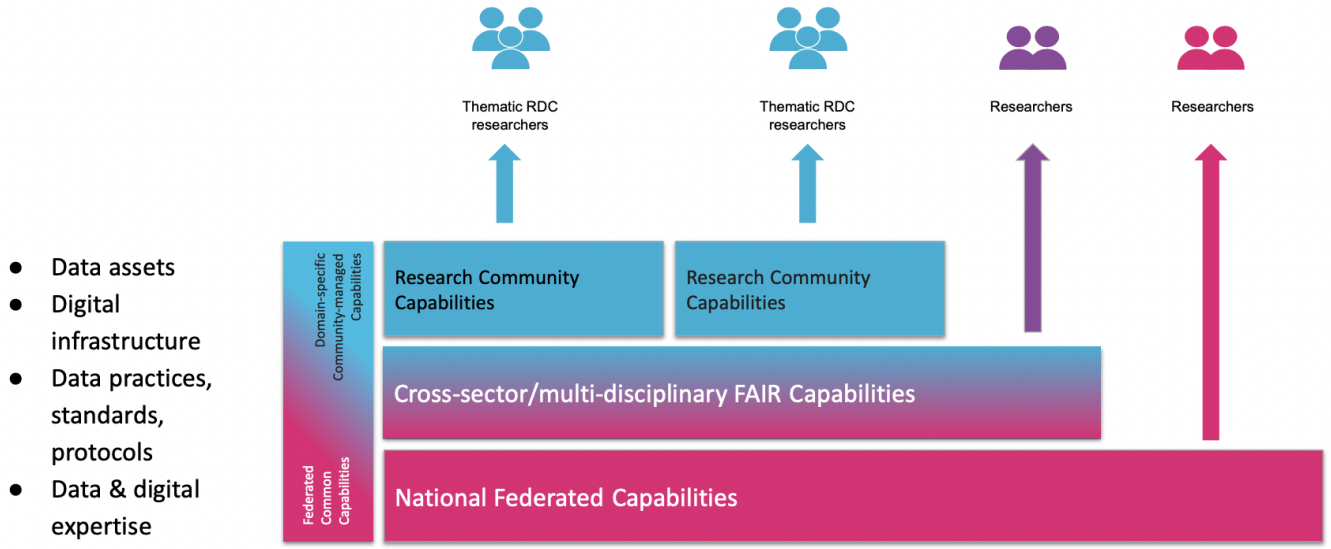

Figure 1: Thematic RDCs drive cross-sector convergence in knowledge infrastructure to support research

The Thematic RDC model will build on the collective capabilities of ARDC, NCRIS, institutional partners and research communities to deliver data assets, digital infrastructure (like storage, compute, services, tools and platforms), data practices, standards and protocols, and data and digital expertise. Importantly, researchers whose research field falls outside the Thematic RDC areas will still have access to underpinning capabilities (Figure 1).

The ARDC is establishing two pilot Thematic RDCs in the 2022-23 financial year with $15.8 million from the 2020 Research Infrastructure Investment Plan to start the pilots. The pilot themes are ‘People’, which has a focus on ‘Health’, and ‘Planet’ which covers ‘Environment and Agriculture’.

‘People’ and ‘Planet’ are working titles for the pilot Thematic RDCs for the purpose of this paper.

DESIGNING THE THEMATIC RDC

The ARDC's nationally focused capabilities in the four portfolios of 'People & Policy', 'Platforms & Software', 'Data & Services' and 'Storage & Compute' that strengthen and support the broader system will focus on identified national challenges and opportunities in the pilot thematic areas of 'Health' and 'Environment & Agriculture' to drive best practice in the creation, analysis and retention of high-quality data assets.

Conceptual Model of a Thematic RDC

Conceptually, the key components of the delivery model for a Thematic RDC are:

- **Expertise** - deep knowledge base within the ARDC team related to data, the research data lifecycle and digital research infrastructure
- **National or Federated Services & Infrastructure** - enduring and underpinning capabilities backed by the ARDC's long-term commitment
- **Projects** - activities that will be undertaken in partnership with national stakeholders to develop new digital research infrastructure

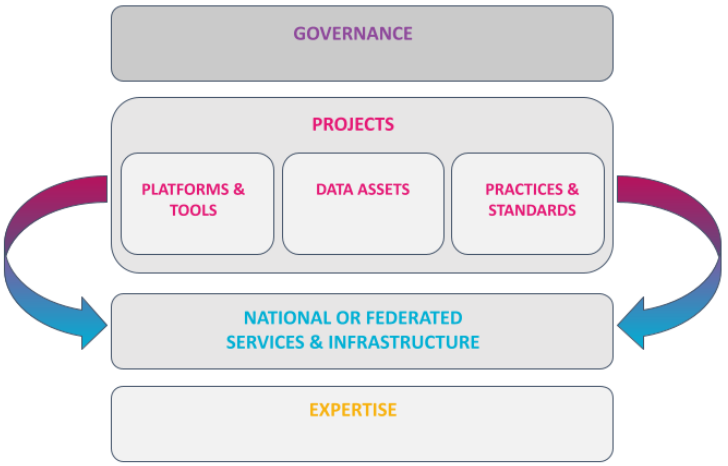

Figure 2: Thematic RDC Conceptual Model

- **Governance** - program governance to ensure the alignment of the Thematic RDC with national priorities and the digital research infrastructure needs of stakeholders across research, government and industry.

PEOPLE RDC - IDENTIFICATION OF CHALLENGE AREAS

How do we know what researchers need for data driven health research?

Analysis of the national priority areas and the research funding priorities related to health research has highlighted potential areas of focus[2]. Further, the Learning Health System Framework[3] developed by the

[2] Appendix: National priority areas and health research funding priorities
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8580135/

Academic Health Research Translation Centres identifies key data and information systems needed to underpin evidence-based health research and translation. The 'Data and Information Systems' requirements outlined in the framework that directly influence the national digital research infrastructure capabilities are:

- Quality, harmonised data from health care and other sources, including patient reported experience and outcome measures
- Compliance with 5 Safes, FAIR data principles and legislative and privacy requirements
- Governance, data sharing, linkage, analysis and interpretation
- Big data analytics, machine learning
- Technology and infrastructure

By investigating these digital research infrastructure requirements and assessing the breadth of projects being currently supported by the ARDC, we can map a pathway to the digital research capabilities that can be delivered by the People RDC.

Many of the data challenges in health research arise from the sensitive nature of health data and jurisdictional and regulatory requirements that lead to a heterogenous digital infrastructure ecosystem. This is manifested in the following data challenges:

- Data discovery across distributed national data assets
- Data collaboration challenges across diverse data sharing and analysis environments
- Application of advanced analytics over distributed data assets
- Data linkage at scale for research and research translation

The ability to deliver consistent practices, technical interoperability and common standards across this diversity will be a defining feature of the People RDC.

These requirements mean that the ARDC is uniquely positioned to deliver the optimal solutions.

PEOPLE RDC - ADDRESSING THE DATA CHALLENGES

Four challenge areas for the People RDC have been identified:

- Provide **federated data discovery** across distributed national data assets
- Support **federated data sharing and analysis** to enable collaborations across diverse environments

- Facilitate **federated analytics for distributed data**
- Accelerate **federated data linkage** to deliver data linkage at scale for research and research translation

The data challenges in the People RDC arise from the diversity across the national health research data ecosystem, ranging from multiple data modalities and data sources through to the cross-sector research data stakeholders, their data requirements and infrastructure for data collaborations.

The following challenges and solutions will be interrogated during the consultation process.

Federated data discovery

Challenge: Enable researchers to find national data assets for reuse outside the initial research/operational activity that created them.

Health research has a rich data ecosystem with national data assets distributed across the government, research and health service sectors. Some examples of these datasets are:

- ➢ Federal: Medical Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS) data from AIHW; and the Multi-Agency Data Integration Project (MADIP) data from ABS
- ➢ State: linked administrative and health data like Admitted Patient data and Ambulance data
- ➢ Research: data from clinical trials, cohort studies and clinical quality registries
- ➢ Health Services: data from hospital electronic medical records (EMR)

The challenge for potential users of these national data assets, like researchers or industry members, is identifying what data is available, where it is located and how it can be accessed.

Proposed solution: Implement collection of metadata from sensitive data collections across diverse national data sources and institutional data repositories in order to populate a discovery service.

Federated data sharing and analysis

Challenge: Enable access and sharing of data extracts from distributed sensitive data collections for national scale data collaborations and analysis; and facilitate researchers and data requesters to work securely within their own institutions.

Data custodians and their organisations typically operate bespoke secure environments for data analysis that enable researchers, collaborators and data requesters to analyse data while ensuring that their national data assets are protected. These institutional secure environments focus on protecting data, but they also lead to data silos that constrain national scale cross-sector, cross-jurisdiction or

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

multidisciplinary data collaborations. When research or research translation requires the aggregation of data from multiple data sources, then it is challenging to work across diverse institutional secure environments.

Additionally from a data requester perspective, enabling researchers and data requesters to analyse shared data extracts within their own institutional secure analysis environment, while maintaining the data governance oversight of the data custodian, removes the need for access to analysis platforms at other institutions.

Proposed solution: Implement interoperability between institutional secure data analysis platforms. This includes privacy-preserving technical interoperability across secure analysis platforms as well as seamless data governance across these platforms to ensure that data remains protected and under the full control of the data custodian at all times.

Federated analytics for distributed data

Challenge: Facilitate the application of advanced analytics techniques like machine learning when national sensitive data collections are managed across disparate institutional secure data repositories, with challenges around data heterogeneity, quality and integration

For example - in the case of federated machine learning, the machine learning model is taken to the data rather than bringing data to the machine learning model. This provides a mechanism to train machine learning models on each of the distributed datasets without having to collate all the data in a central repository.

Proposed solution: Where data integration is not feasible, evolve frameworks, best practices, tools and expertise for distributed analytics underpinned by fit-for-purpose digital research infrastructure.

Federated data linkage

Challenge: Accelerating and scaling federated data linkage capabilities to meet research demand

Data linkage brings together data from different sources relating to the same individual or event.

Australia has world-leading, NCRIS funded, data linkage capabilities which enable a wide range of data to be linked within and across jurisdictions. This national data linkage infrastructure is used to link population level administrative data e.g. hospital admissions, Pharmaceutical Benefits Scheme as well as to link research data to state/territory and Australian Government administrative data.

Proposed solution: Building on existing national data linkage infrastructure, propagate and develop frameworks, best practices, tools and expertise to scale up data linkage capabilities nationally, while ensuring the highest standards of privacy, security and quality.

APPENDIX

People RDC national priority areas and health research funding priorities

National priority areas

The national priorities related to health research and supporting infrastructure are addressed in the following national strategies:

- 2021 National Research Infrastructure Roadmap Exposure Draft[4]
- National Medical Research & Innovation Strategy[5]
- National Science & Research Priorities[6]
- National Climate Resilience & Adaptation Strategy[7]
- Blueprint for Critical Technologies[8] and The Action Plan for Critical Technologies[9]

Looking across these high-level strategic documents, a number of areas of focus can be identified. These are summarised below:

- Development of biologics and medical devices
- Pre-clinical and clinical research
- Diverse national health datasets and integrated digital platforms
- Enabling research and research translation to support better models of health care and services for all sections of the community
- Cutting-edge treatments with genomics and genetic engineering

Health research funding priorities

National health research is largely funded through NHMRC and MRFF.

[4] https://www.dese.gov.au/national-research-infrastructure/resources/2021-national-research-infrastructure-roadmap-exposure-draft
[5] https://www.health.gov.au/sites/default/files/documents/2021/11/australian-medical-research-and-innovation-strategy-2021-2026.pdf
[6] https://www.industry.gov.au/sites/default/files/2018-10/science_and_research_priorities_2015.pdf?acsf_files_redirect
[7] https://www.awe.gov.au/sites/default/files/documents/national-climate-resilience-and-adaptation-strategy.pdf
[8] https://www.pmc.gov.au/sites/default/files/publications/ctpco-blueprint-critical-technology.pdf
[9] https://www.pmc.gov.au/sites/default/files/publications/ctpco-action-plan-for-critical-technology-amalgamated.pdf

The NHMRC research priorities[10] are dementia, mental health and health impacts of environmental change.

MRFF research Priorities outlined in the Australian Medical Research and Innovation Priorities 2021-26[11] highlight the importance of:

- Research and research translation that is multidisciplinary, cross-sector and cross-jurisdiction on a national scale
- Enabling research and research translation in primary care settings and clinicians in the health care system; and indigenous researchers
- Bringing together diverse datasets from priority populations - through data linkage along with data storage and analytics
- Data platforms, Applied AI, novel decision tools and end-user digital utility

[10] https://www.nhmrc.gov.au/research-policy/research-priorities
[11] https://www.health.gov.au/sites/default/files/documents/2021/11/draft-australian-medical-research-and-innovation-priorities.pdf