Australian Research Data Commons

Planet Research Data Commons

PROGRAM DESCRIPTION

30 June 2023

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

The ARDC is enabled by NCRIS

CONTENTS

DOI: 10.5281/zenodo.8098119

Suggested citation:

ARDC Ltd. (2023). *Planet Research Data Commons Program Description*.
Viewed online at:
https://doi.org/10.5281/zenodo.8098119

INTRODUCTION

The Australian Research Data Commons (ARDC) has adopted a Thematic Research Data Commons (RDC) approach as a vehicle for the ARDC and our national partners to collaboratively develop and deliver sustainable digital research infrastructure on a national scale. The Thematic RDCs are enduring initiatives, with the digital research infrastructure capabilities being sustained and enhanced over the long-term. They enable us to best meet the needs of our diverse national research communities in a strategic and comprehensive way.

The *Planet Research Data Commons* (Planet RDC) for environmental and earth science research is one of 3 Thematic RDCs launched by ARDC.

The Planet RDC supports environmental and earth science researchers and decision makers to develop cross-sector and multi-disciplinary data collaborations on a national scale. It will enable trusted data and information supply chains so that research, government and industry users can seamlessly access trusted data and models to inform management decisions, and deliver unprecedented insights into current and future states of the environment.

This program description outlines the drivers behind establishing the Planet RDC and the national consultation process. It then describes the vision, the high-level program strategy, and activities that the Planet RDC will undertake *during establishment*. The program description will be used to inform collaborations, programs and projects undertaken by the Planet RDC.

NATIONAL CHALLENGES

Research Drivers

Earth and environmental science seeks to address some of the most complex, interconnected and integrated challenges facing society. These include responding to a changing climate, increasing agricultural productivity, biosecurity risks, accommodating a growing human population and related land use changes, transitioning to new energy systems, managing geohazards, declining biodiversity, and understanding the importance of marine, coastal and freshwater environments for ecosystem services.

Increasingly, research is focused on understanding broad domains, and at the interface between domains The challenges include:

- The complexity of environment and earth science domains is too high for any single institution to address alone.
- Integrating data requires a large effort.
- Research technology platforms are difficult to sustain and operate beyond the duration of grant funding or achieve critical mass for sustainability.
- Does not capture the benefits of innovation for reuse across the sector.

There is the need for a deep and sustained research infrastructure platform that increases connectivity, interoperability and standardisation of data across domains and scales.

Policy Drivers

Multiple national science and regulatory policies have called for improved connectivity, availability and quality of data systems and analysis tools to support research and decision making. The *2021 State of Environment Report* highlights the necessity of coordinated data infrastructure across government, industry and national research infrastructures to effectively adapt to environmental and climate-related changes. Samuel's *Independent Review of the EPBC Act* emphasises the need to understand cumulative impacts of developments in a region, and highlights the rate of environmental decline.

The *Royal Commission into National Natural Disaster Arrangements* (the Bushfires Royal Commission) and *Nature Positive Plan: Better for the Environment, Business for Business* both emphasise the need for better data quality, standardisation and integration. These reviews recommended the development of data sharing agreements and protocols between government agencies, research institutions, and the private sector to ensure that data is accessible and usable.

Evolving policy and regulatory positions, particularly in relation to sustainability and renewable energy, highlight the need for increased data availability and connectivity across multiple domains and sectors. This encompasses research domains such as marine science, ecology, biology, urban planning and the geosciences, as well as government, industry and research institutions.

New methods to manage the environment such as market-based approaches and natural capital accounting will require reliable data and information supply chains and tools that are auditable and reproducible. Both government and industry are significant contributors and users of data and models. However, challenges persist in accessing sensitive data, addressing commercial considerations, translation of research into practice and establishing appropriate national standards to facilitate integration and access.

Digital Drivers

Australia possesses world-class environmental research infrastructure supported by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS). This infrastructure supports the monitoring, collection and management of data across a variety of domains and disciplines.

The *2021 National Research Infrastructure Roadmap* highlights the critical need for national digital research infrastructure capable of supporting seamless multidisciplinary research on a national-scale. Researchers need digital infrastructure that enables the sharing and reuse of earth and environmental data in an interoperable, machine readable and understandable form.[1]

Digital technologies have been instrumental in monitoring and understanding the environment. With the advent of sensors, the Internet of Things (IoT), remote sensing, and new data collection approaches like citizen science, both the quantity and diversity of environmental data have substantially grown. This data, as with data from more traditional sources, is often heterogeneous and fragmented.

To effectively manage the escalating volumes of data and extract meaningful insights, new data science methods such as artificial intelligence and machine learning, as well as cloud-based technologies for data integration and storage, are essential. Furthermore, ensuring data availability in multi-cloud environments allows researchers to leverage diverse technologies and resources.

To advance our understanding of the environment, it is crucial to have access to integrated, scalable, and dynamic data across all domains, along with interoperable and integrated models. These are key components in developing next-generation dynamic models and digital environmental twins. This needs to be supported by advanced identity and access controls, robust cyber security practices and seamless connectivity. Researchers, data scientists, decision-makers, and computer science experts need to be able to collaborate to advance our understanding of environmental changes.

[1] Zhai, Z., Hellstrom M. (2022), Towards Interoperable Research Infrastructure for Environment and Earth Sciences. A reference model guided approach to common challenges. *Springer Nature Switzerland.* https://doi.org/10.1007/978-3-030-52829-4

National Consultation

To determine the scope, scale, and focus of the digital capabilities to be delivered through the Planet RDC, the ARDC conducted an extensive *national consultation* and community engagement. This included:

- A national consultation process consisting of roundtables and targeted consultation with over 130 participants.
- Workshop and meetings with directors of National Earth and Environment Science Facilities (NEESF).
- Workshop and meeting with all Planet-relevant Platform and National Data Asset programs.
- Workshop on Trusted Data and Information Supply Chains with industry, government and research sectors.

Data challenges, technological barriers, and opportunities were examined and prioritised that the Planet RDC could address to make a significant impact at the national level. The *key outcomes* of these consultations are summarised in Figure 1 below.

The digital infrastructure needs of the earth and environmental sciences community were collated under four data challenge areas: Reliable and trusted data and information supply chains; Curated, integrated datasets and services; Modelling, analytics and decision support infrastructure; and Indigenous knowledge management, FAIR and Skills.

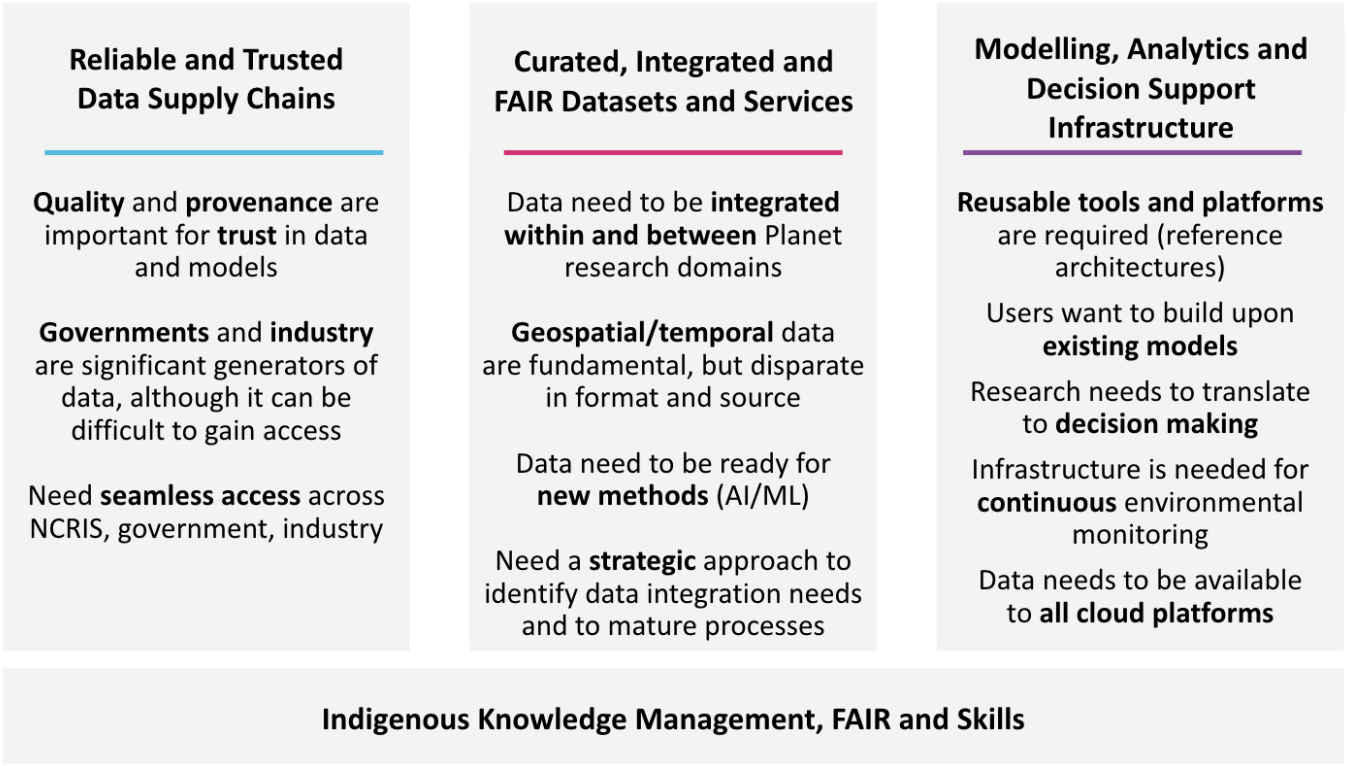


Figure 1. Outcomes of the national consultation: digital infrastructure needs of the earth and environmental science communities

THE PLANET RDC VISION

A DYNAMIC, INTERCONNECTED AND INTEGRATED DIGITAL FUTURE

The Planet RDC envisages a future where research, government and industry can seamlessly access, visualise, analyse and model trusted data that is integrated across earth and environmental domains, time and spatial scales.

It charts a vision where research communities are able to network models and their outputs together across different disciplines such as climate, environmental, marine, and terrestrial sciences. This integration enables groundbreaking research, informs decision-making and delivers unprecedented insights into current and future states of the environment. This follows other international initiatives such as the *Destination Earth* (EU) and the *NERC Digital Strategy 2021 to 2030* (UK).

The Planet RDC will provide infrastructure and coordination to enable research communities to take advantage of new data science methods, advances in cloud technology, networked modelling and integrated data to further accelerate research and impact.

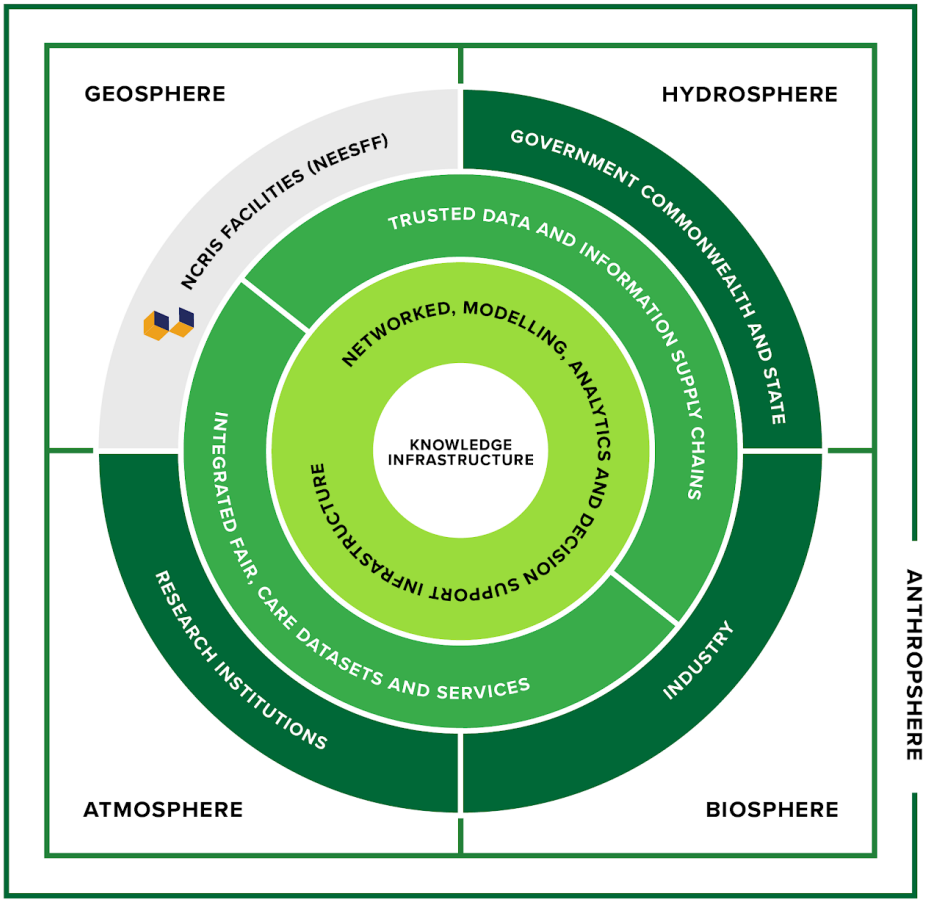


Figure 2. The Planet RDC is working collaboratively towards a networked earth and environmental infrastructure

It will forge national partnerships between research, government and industry to facilitate the translation of research findings and methodologies into actionable solutions to address national priorities.

The Planet RDC will work with all partners to incorporate Indigenous knowledge management, including the CARE principles, into national research infrastructures.

By delivering these elements, Australia's capacity to understand, predict and manage environmental changes will be significantly enhanced.

HOW WILL WE DO THIS?

Planet RDC Program Strategy

The Planet RDC will apply ARDC capabilities and expertise in digital research infrastructure to national challenges in the earth and environmental science domains. The Planet RDC will focus on the four data challenge areas, delivering interconnected and enduring digital infrastructure and establishing strong research translation pathways between research, government and industry. This will be delivered as a set of complementary activities that align with and extend the capabilities of our partners in National Research Infrastructures, government and industry.

To achieve this, the ARDC will collaborate with Australian organisations, including the public and private sectors, research institutions (such as universities and publicly funded research agencies), research infrastructure providers and other relevant stakeholders. Together, we will co-design and implement digital research infrastructure within the Planet RDC through collaborative partnerships.

The Planet RDC will focus on culture change, working with institutions and government to ensure appropriate incentives are in place to enhance data sharing and interoperability, and fostering trust between stakeholders.

The infrastructure and services developed will be operationalised, allowing the community to use and build upon them with confidence. The capabilities to be developed will be determined through consultation with the community. They may include: operational responsibilities for core platform infrastructure and services such as a managed Kubernetes service, or infrastructure for data-as-a-service, environmental data spaces, meshes and lake storage. Additionally, machine observation/sensor processing services, and a repository for national reference data assets (including model outputs) will be considered.

The four data challenge areas will be integrated through a focus on overarching use cases based on the data infrastructure needs of government, research and industry in selected regions of national interest and priority.

1. Trusted Data and Information Supply Chains

The need for trusted data and information supply chains is paramount across research, government, and industry sectors. Researchers require accurate and dependable data to conduct their studies and derive valid conclusions. Governments need reliable data for policy-making, environmental management, and regulatory compliance. Industries require trustworthy data for environmental impact assessments, sustainable resource management, and to meet regulatory standards.

Australia has many excellent producers of data. However, there is currently no whole-of-system approach to overcome the challenges of discovering and accessing data, the costliness of making it interoperable, and its limited scalability. A system-wide approach can be achieved through:

- Supporting a common view of the capabilities in earth and environmental science data and information supply chains, enabling better alignment of effort and interoperability.
- Increasing trust mechanisms across the earth and environmental sectors to facilitate increased data flows between research, government and industry.

Trust needs to be enhanced across the entire information supply chain. This includes data but also the models and analytics that underpin and develop new knowledge and knowledge products. As with data, to ensure reliability and reproducibility, models and their outputs need to be FAIR (findable, accessible, interoperable, and reusable).

The transaction costs of accessing and sharing data will be lowered through the adoption of national data standards, system interoperability, infrastructure accreditation, and standard governance and technical frameworks for data exchange such as International Data Spaces. This would result in productivity gains, new research and increased robustness of information and knowledge generation.

The program will work closely with government and industry-led initiatives and partner with international initiatives such as *CODATA* and the *Research Data Alliance* to facilitate the transfer of knowledge and shared learning.

How will we do this?

1. Support a common view of the capabilities in earth and environmental science data and information supply chains, enabling better alignment of effort and interoperability. This will be achieved through the use of the *Shared Analytic Framework for the Environment (SAFE 2.0)* (see *Figure 3*).
2. Identify, influence and implement mechanisms required for the data and information supply chain to operate effectively. This will include mechanisms for accrediting the trustworthiness of providers of data and services (e.g *CoreTrustSeal*), common data governance frameworks, data sharing protocols, licences and for operationalising agreements (e.g. service-level agreements) between organisations for providing enduring data and services.

3. Identify means to reduce the cost of data integration, including data being FAIR from the point of creation and throughout the supply chain.
4. Deliver interoperability uplift across the NRI, government and research sectors, by supporting the implementation of community-agreed data and metadata standards, vocabularies/ontologies, identifiers, and best practices for data quality and provenance.
5. Deliver exemplar data and information supply chain projects that combine research, government and industry use cases across multiple domains and disciplines.

2. Integrated FAIR Datasets and Services

Being able to understand, predict and manage changes to the environment requires continental scale observations backed by a strong observation network that is integrated and interoperable.

However, finding, accessing and standardising data for cross domain analysis is still resource intensive. Data is often too disparate and heterogeneous to be easily interoperable. The provision of a single entry point to FAIR data, alongside next generation storage and repository services is needed to realise the Planet RDC vision.

Research institutions produce high-impact but long-tail data across many disciplines. Managing the large number of heterogeneous datasets is challenging. However, it is critical to enable the linking and networking of significant data collection and research effort at individual, institutional and regional scales. The focus needs to be on equipping data stewards with the tools and capability to enable long-tail data to be made FAIR, to enable integration for knowledge creation.

New data-as-a-service and contemporary cloud based data infrastructures (such as data lakes and data meshes) need to be implemented to enable future data science capabilities. These services will be able to assimilate data from multiple sources, and make it available to multiple platforms and functions. Data and services need to be ready for new data science methods such as AI/ML which is able to be implemented over different data types, scales and volumes.

How will we do this?

1. Provide national leadership and work with research, national research infrastructures and government partners to strategically prioritise data integration, national standardisation and scientific model development projects. This will ensure Planet RDC project outputs meet national research challenges, and enable research translation into policy, management and operations.
2. Support the mobilisation, accessibility and interoperability of data from research, NRI, government and industry. The data assets will be made FAIR and analysis-ready to ensure they are available for new analysis methods such as AI/ML. This includes the use of ARDC's National Information Infrastructure such as persistent identifiers and Research Vocabularies Australia.

3. Provide expertise to major research communities such as Centres of Excellence and National Environmental Science Program (NESP) hubs to make their high-impact, long-tail data FAIR and analysis-ready.
4. Enable federated discovery of, and access to, 'Federated Data Collections' of discipline or facility datasets and software. This will be achieved through system enhancements and processes for presenting datasets harvested from institutional and government repositories as Themed Collections in Research Data Australia.
5. Provide data storage and cataloguing services for curated, integrated FAIR datasets that can be used by multiple platforms and services (including Networked Modelling, Analytics and Decision Support Infrastructure).
6. Work with research communities via the National Earth and Environmental Facilities Forum (NEESFF) to prioritise and develop *FAIR implementation profiles (FIPs)* relevant to each domain / discipline.

3. Networked Modelling, Analytics and Decision Support Infrastructure

Virtual research environments, research platforms and data visualisation tools empower researchers to collaborate with national and international partners in modelling and data analysis. They provide streamlined access to data, advanced analytic tools and compute resources, and encourage repeatable workflows. Platforms that combine user interfaces with trusted models and data enable research and its translation into decision making and management.

However, differing technical architectures restricts interoperability and integration between platforms, hindering multidisciplinary research. Additionally, these research platforms are often developed using project or grant funding and ongoing sustainability is a challenge, as is access to appropriately skilled technical staff.

National coordination is required to realise economies of scale in staff resources and skills. Coordination of shared digital infrastructure and reference architectures will significantly enhance the capability of research communities to deploy modelling and analytic platforms at scale.

Through alignment and reuse of infrastructure, research communities will be able to focus on developing new scientific models and gaining new insights from them, rather than on information technology infrastructure and maintenance.

Support is also needed to enable the reuse of scientific models, and the ability to network between models to enable multidisciplinary research. Models and their outputs need to be treated as fundamental assets with services, catalogues and technical staff supporting their curation.

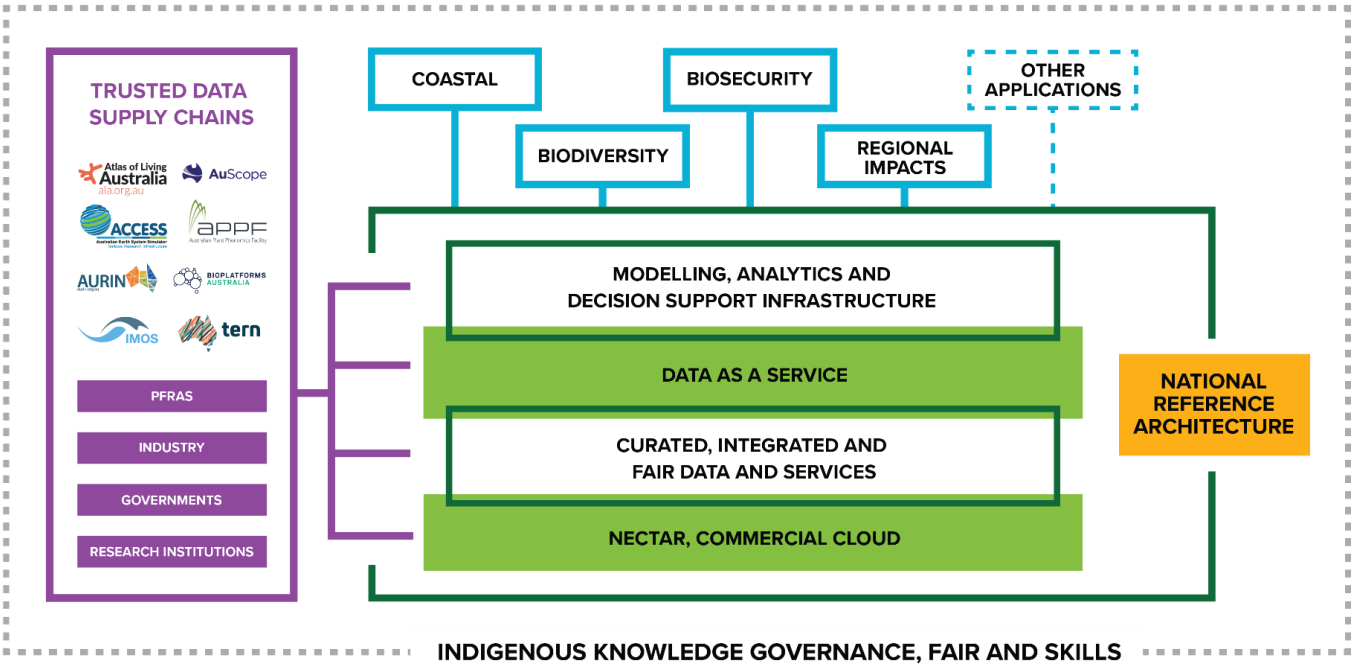


Figure 2. Planet RDC national reference architectures will allow research communities to integrate data, models and analytics at scale

National environmental data should be available in multi-cloud environments allowing researchers to access the most appropriate modelling tools and analytic frameworks.

Provision of enduring infrastructure will allow research, government and industry to build trust and confidence in digital research infrastructure, and the ability to enact long-term investment.

How will we do this?

1. Develop a reference architecture(s) for the Planet RDC digital infrastructure to enable research communities to easily build and deploy platforms and tools into a cohesive Planet RDC ecosystem at scale.
2. Develop and maintain extensible underpinning platform-as-a-service infrastructure that can be used by modelling research communities. This will enable faster and more efficient deployment of cloud-based modelling, analytics (e.g R, python, AI/ ML workflows) and decision-support capabilities (e.g. decision tools, data visualisation).
3. Provide shared infrastructure and services for the sharing of scientific models and their outputs, to ensure they are available for reuse and integrating with models from other disciplines.
4. Support the establishment of national shared infrastructure and services to enable processing of machine observation data (e.g. imagery, acoustics, video). This may include working storage and compute infrastructure, shared processing scripts, and the implementation of persistent identifiers for observation data and associated instrumentation.

5. Enable seamless discovery of, and access to, federated authoritative spatial data, modelled outputs and time-series data across Planet domains. This will be achieved through the provision of infrastructure and supporting the implementation of standards.
6. Enable seamless access to sensitive data across sectors by implementing consistent national governance and authentication and authorisation protocols, and providing access to systems that enact these protocols.
7. Develop national partnerships with commercial cloud services that enable multi-cloud data provision and analysis environments.

4. Indigenous knowledge management, FAIR and Skills

"Indigenous knowledge and sustainable cultural practice are key to environmental management" (*2021 SOE 2021 Key findings*).

Indigenous knowledge is essential to caring for Country (land, sea and water), and therefore to protecting and restoring Australia's natural environment. The Planet RDC will work with our NCRIS partners to implement a common best practice approach to Indigenous knowledge management. For example, implementing the CARE principles (Collective Benefit; Authority to Control; Responsibility and Ethics) for Indigenous data governance, as these were endorsed in our consultations as foundational for the Planet RDC.

Making data and research software findable, accessible, interoperable and reusable (FAIR) is recognised as a key requirement that underpins the other data challenges. Multi-disciplinary research into complex environmental challenges requires data from different sources and different domains to be discovered, easily accessed and integrated. Data supply chains can flow more freely when data are born-FAIR, and it is essential that clear provenance information is available for trust data and software. The ARDC, via the Planet RDC, will continue to engage with and influence stakeholders to improve the culture of FAIR data and break down intra- & cross-sectoral silos.

How will we do this?

1. In collaboration with the HASS RDC and Indigenous Research Capability and other NRI, develop and enact CARE implementation guidelines to support policy and practice change.
2. Develop and implement skills and training frameworks for earth and environmental science researchers and practitioners to ensure a digitally skilled workforce.
3. Deliver an engagement program to promote best practices for FAIR earth and environmental data and software in Government and Industry sectors.

SAFE – Layers and Capabilities

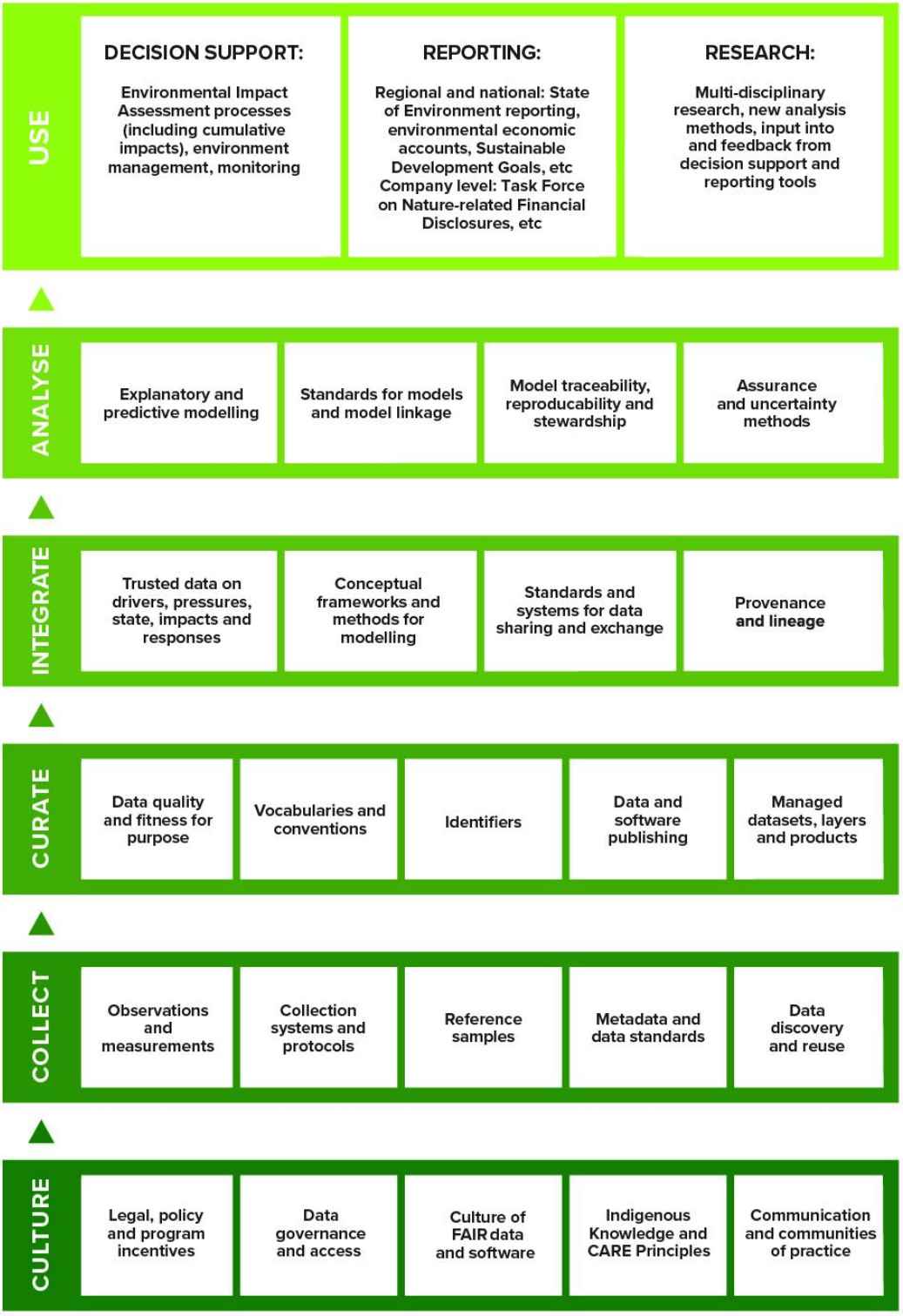
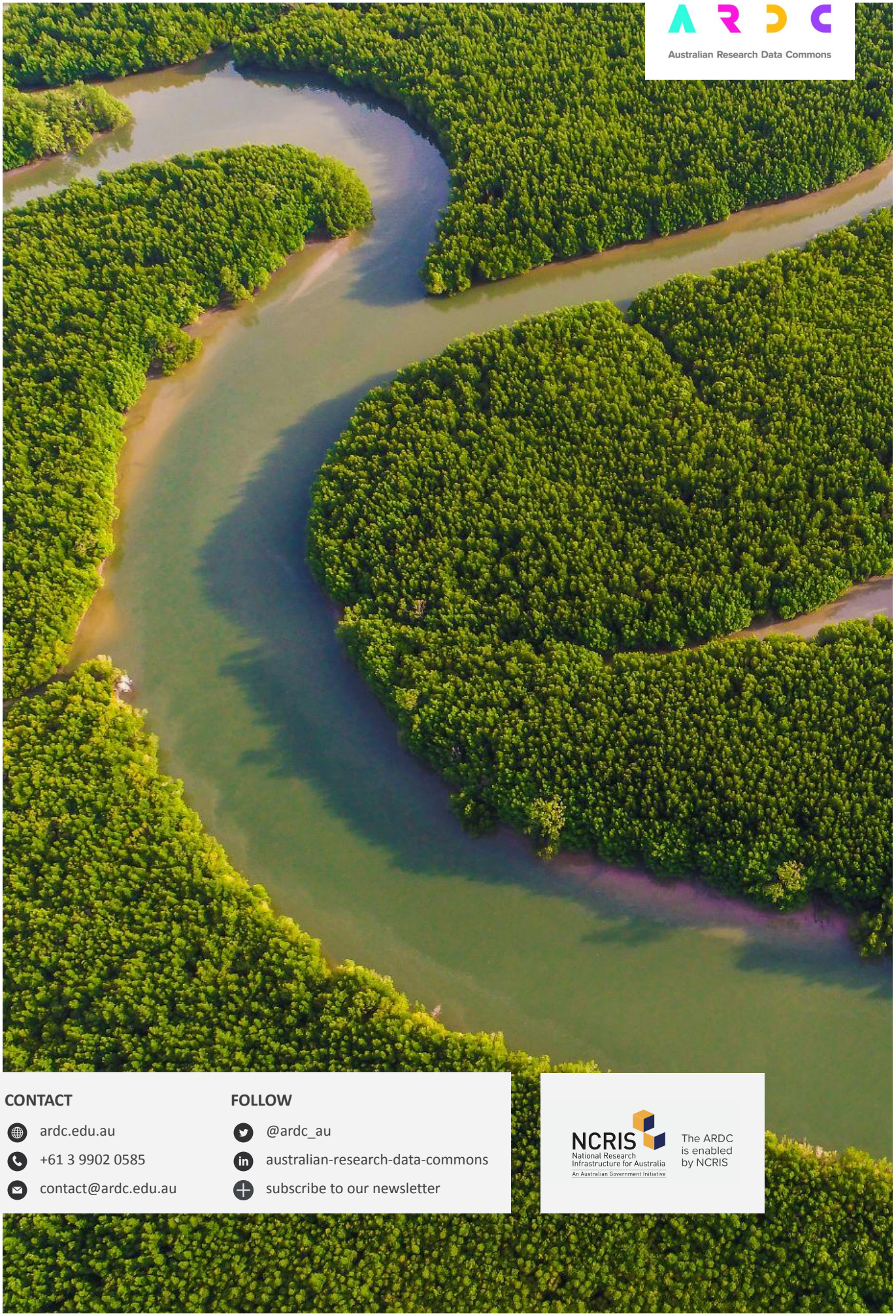


Figure 3. The Shared Analytic Framework for the Environment (SAFE 2.0) presents a common view of the capabilities in earth and environmental science data and information supply chains. This framework will be utilised across the Planet RDC to ensure alignment and interoperability. (WABSI, WAMSI and ARDC 2023, Capabilities for a national supply chain of environmental information SAFE 2.0. The Western Australian Biodiversity Science Institute.)



Australian Research Data Commons

CONTACT

ardc.edu.au

+61 3 9902 0585

contact@ardc.edu.au

FOLLOW

@ardc_au

australian-research-data-commons

subscribe to our newsletter

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

The ARDC is enabled by NCRIS