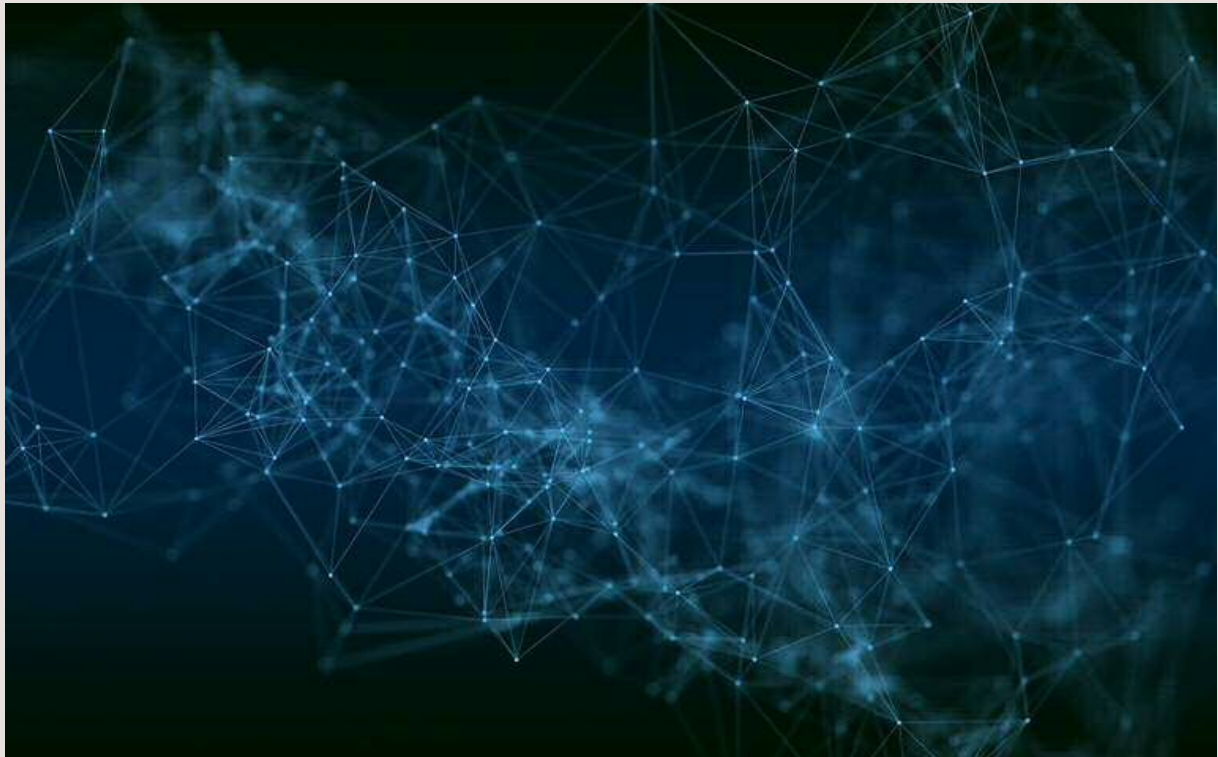


The Research Data Landscape in Aotearoa New Zealand

A report undertaken in partnership with the Aotearoa New Zealand Committee on Data in Research (CoDiR).



DOI: <https://doi.org/10.15663/UoW.RDLA.DEC2023>

CONTRIBUTORS

Rogena Sterling, The University of Waikato | Te Whare Wānanga o Waikato

 [0000-0002-5949-387X](https://orcid.org/0000-0002-5949-387X)

Michelle Blake, The University of Waikato | Te Whare Wānanga o Waikato

 [0000-0002-3647-9848](https://orcid.org/0000-0002-3647-9848)

Nick Jones, University of Auckland | Waipapa Taumata Rau

 [0000-0001-5513-8312](https://orcid.org/0000-0001-5513-8312)

Richard Hartshorn, University of Canterbury | Te Whare Wānanga o Waitaha

 [0000-0002-6737-6200](https://orcid.org/0000-0002-6737-6200)

Tahu Kukutai, The University of Waikato | Te Whare Wānanga o Waikato

 [0000-0001-5080-2296](https://orcid.org/0000-0001-5080-2296)

ACKNOWLEDGEMENTS

This report, undertaken by a project group from the Committee on Data in Research (CoDiR), was commissioned by the Ministry of Business, Innovation and Employment (MBIE).

Suggested citation:

Sterling, R., Blake, M., Jones, N., Hartshorn, R., & Kukutai, T. (2023). *The Research Data Landscape in Aotearoa New Zealand: A report undertaken in partnership with the Aotearoa New Zealand Committee on Data in Research (CoDiR)*. Ministry of Business, Innovation and Employment. <https://doi.org/10.15663/UoW.RDLA.DEC2023>



**MINISTRY OF BUSINESS,
INNOVATION & EMPLOYMENT**
HĪKINA WHAKATUTUKI

EXECUTIVE SUMMARY

Data and datasets are often described as a core strategic asset for Aotearoa New Zealand (Aotearoa) and indispensable for the government's ambition of being a small nation with an advanced, adaptive, and inclusive economy. In the context of our research, science and innovation (RSI) ecosystem, the value of data cannot be overstated.

The purpose of this report is to provide a review (the Review) of Aotearoa's research data landscape. Research data are data that are used as primary sources to support technical or scientific enquiry, research, or artistic activity; as evidence in the research process; and/or are commonly accepted in the research community as necessary to provide a foundation for, or validate research findings and results. The Review covers four core areas:

- Te Tiriti o Waitangi and data sovereignty
- research data ecosystems
- research data infrastructure
- research data cultures.

Although informed by international data landscape reviews, this Review emphasises the unique considerations and structural features of the Aotearoa data landscape. Based on our analysis and observations from a series of targeted workshops, we provide a set of recommendations on how to strengthen the system and advance shared aspirations for better outcomes. The recommendations are organised under the five headings from Nosek's Pyramid of Social Change, setting out a phased strategy for culture and behaviour change. In implementing these recommendations, we recognise that the articles of Te Tiriti should be embedded throughout, consistent with sector requirements (MBIE, 2023d).

RECOMMENDATIONS (ABRIDGED)

INFRASTRUCTURES – MAKE IT POSSIBLE

1. Clarify which national data repositories are available for researchers in Aotearoa to deposit their data. Special attention should be given to locally owned sovereign repositories that can safeguard Māori research data within Aotearoa.
2. Support a review of current research data repositories and databases to assess how research organisations can implement recommendations from the Māori Data Governance Model, and identify further work required.
3. Develop a bicultural data ontology, building on existing work.
4. Improve support for domain- and community-specific infrastructures, including identifying next steps related to the review of nationally significant databases and collections.

5. Support MBIE's work to improve national eResearch infrastructure.

USER INTERFACES/EXPERIENCES – MAKE IT EASY

6. Support a community-led effort to examine national and international experiences on RDM across different disciplines/domains.
7. Provide practical resources to support improving the design of RDM services and engagement with stakeholders across a range of disciplines, and help to provide practical guidance to researchers in achieving best practice.

CULTURES: COMMUNITY / BEHAVIOUR / NORMS – MAKE IT NORMATIVE

8. Support the development and implementation of a national strategy for open research, incorporating the FAIR and CARE Principles.
9. Establish a cross-sector community of practice in RDM tasked with fostering a culture of collaboration among researchers, institutions and the wider community.
10. Promote community engagement in the research process, including participatory governance of data.
11. Develop training and education initiatives to enhance research data culture and promote data sharing.
12. Support community efforts to work towards common approaches, processes and guidelines to shape research data culture, including ethical data use and privacy protection.

INCENTIVES – MAKE IT REWARDING

13. Support open research goals (implementing the FAIR and CARE Principles).
14. Implement incentives that recognise and reward researchers for adhering to data sharing and management practices.
15. Support and encourage ways of celebrating best practice.

POLICIES – MAKE IT REQUIRED

16. Move towards mandating RDM as a requirement of researcher funding once other essential elements are in place above, such as infrastructure and training.
17. Māori governance over Māori research data should be socialised with research organisations and formalised as a requirement of research funding.

18. Develop and maintain data policies that adhere to the FAIR and CARE Principles, ensuring accessibility and usability of research data and MDGov compliance.
19. Set clear expectations for ethical data use and privacy protection through policy implementation.
20. Ensure policies contain guidance that outlines standards for RDM, including documentation, storage and preservation.

These recommendations provide a comprehensive framework for shaping government policies and initiatives related to research data in Aotearoa.

INITIALISMS

AI	artificial intelligence
AIATSIS	Australian Institute of Aboriginal and Torres Strait Islander Studies
AWS	Amazon Web Services
BC	Biocultural (labels)
CANZUS	Canada, Australia, New Zealand, United States
CARE	C ollective Benefit, A uthority to Control, R esponsibility and E thics (Principles for Indigenous data)
CoDIR	Committee on Data in Research
CONZUL	Council of New Zealand University Libraries
CRIs	Crown Research Institutes
CURFs	confidentialised unit record files
DGMT	Data Governance & Management Toolkit
DLM	data lifecycle management
DM	data management
DMM	data maturity model
DMMM	data management maturity model
DMP	data management plan
DOI	digital object identifier
DPMC	Department of the Prime Minister and Cabinet
ECR	early career researcher
ECMWF	European Centre for Medium-Range Weather Forecasts
FAIR	F indable, A ccessible, I nteroperable and R eusable (Principles for data)
FDBS	federated database system
GIDA	Global Indigenous Data Alliance
GORC	Global Open Research Commons Interest Group
HPC	high-performance computing
IDGov	Indigenous data governance
IDSov	Indigenous data sovereignty

IDI	Integrated Data Infrastructure (a research database)
IEEE	Institute of Electrical and Electronics Engineers
IP	intellectual property
IT	information technology
LBD	Longitudinal Business Database
LLMs	large language models (a type of AI)
MBIE	Ministry of Business, Innovation and Employment
MDGov	Māori data governance
MDSov	Māori data sovereignty
ML	machine learning
NASA	National Aeronautics and Space Administration
NCBI	National Center for Biotechnology Information
NeSI	New Zealand eScience Infrastructure
NIWA	National Institute of Water and Atmospheric Research
NLM	National Library of Medicine (part of the NCBI)
NRPs	national research priorities
NSC	National Science Challenges
NZGOAL	New Zealand Government Open Access and Licensing (Framework)
NZRIS	New Zealand Research Information System
OA	open access
ORCID	open researcher and contributor ID
PBRF	Performance-Based Research Fund
PFGHI	Policy Framework for Government-held Information
PIDs	persistent identifiers
POLDER	Polar Data Discovery Enhancement Research
POSI	Principles of Open Scholarly Infrastructure
PSI	public sector information
RDA IG	Research Data Alliance International Indigenous Data Sovereignty Interest Group
RDM	research data management

REANNZ	Research and Education Advanced Network New Zealand
ROR	Research Organization Registry
RSI	research, science and innovation
SGIG DSC	Self-Governing Indigenous Governments Data Steering Committee
TAP	Te Ara Paerangi (Future Pathways)
TK	Traditional Knowledge (labels)
TMR	Te Mana Raraunga
TRUST	Transparency, Responsibility, User focus, Sustainability and Technology (Principles for digital repositories)
UNESCO	United Nations Educational, Scientific and Cultural Organization

GLOSSARY

hapū	subtribe
iwi	tribe
kaitiaki	guardian, typically of an environmental area or resource
kaitiakitanga	guardianship
karakia	ritual chant, blessing
koawa	waterways, rivers and streams
kōrero	discussion
kotahitanga	unity
kūmara	sweet potato
mana motuhake	Māori self-determination
manaakitanga	reciprocity
mātauranga	Māori knowledge system
mauri	lifeforce
mōteatea	lament, traditional chant
pou	pillard
rongoā	traditional medicinal practices
taonga	treasure; something of particular cultural or spiritual significance
taonga katoa	all the treasured things
taonga tuku iho	heirloom
te ao Māori	the Māori world
tikanga	customary system of values and practices
tino rangatiratanga	full authority
tirohanga	perspective
tuna	eel
waka	canoe
wānanga	forum for deliberations
whakapapa	genealogy, relationships

whānau	(extended) family group
whanaungatanga	sense of family connection developed through kinship rights and obligations
wharehui	meeting house

Contents

Executive Summary	2
Recommendations (abridged)	2
Infrastructures – Make it possible	2
User interfaces/experiences – Make it easy.....	3
Cultures: Community / behaviour / norms – Make it normative.....	3
Incentives – Make it rewarding	3
Policies – Make it required	3
Initialisms	5
Glossary	8
1. Introduction	13
Purpose.....	13
Methodology.....	14
International reviews.....	15
Canada	15
United Kingdom	16
Australia	16
Europe.....	17
United States	17
Global Open Research Commons	18
Aotearoa reviews	19
2. Te Tiriti o Waitangi and Data Sovereignty	21
Te Tiriti o Waitangi	21
Mātauranga	21
Māori data is a taonga	22
Māori data sovereignty.....	23
Māori data governance.....	24
Indigenous data sovereignty and dat Governance.....	24
Working with local and international models of IDsov and IDGov	26
3. Research Data Ecosystems	27
Data ecosystems.....	27
Research data management.....	28
Data discovery.....	30
Data on research projects.....	32
Metadata	33
DOIs	33

PIDs for institutions/projects.....	33
Metadata challenges for Māori	34
Local Contexts.....	34
Open data and open research	35
Open data	35
Open research	35
FAIR and CARE Principles.....	36
4. Data Infrastructure and Systems	37
Data infrastructure	37
Storage.....	38
Data sovereignty and jurisdiction.....	39
Computing systems/power	39
Data repositories, archives and commons	40
Artificial intelligence and research data.....	42
Connective tissue	44
Federated systems.....	45
5. Research Data Cultures	46
Research data cultures.....	46
Benefit-sharing	46
Policies and incentives.....	46
Expectations, norms and cultures	47
Engaging communities	48
Community engagement	48
Citizen science	48
People, skills, expertise and workforce.....	49
Skills/expertise and skill learning.....	49
Workforce	50
6. The Way Forward	52
Summary.....	52
Recommendations.....	53
Infrastructures – Make it possible	53
User interfaces/experiences – Make it easy.....	54
Cultures: Community / behaviour / norms – Make it normative.....	55
Incentives – Make it rewarding	55
Policies – Make it required	56
Bibliography	57
Appendix One: Research Data Landscape Project Proposal.....	66

1. INTRODUCTION

PURPOSE

The New Zealand Government invests over \$2.1 billion in publicly funded research every year with the aim of creating value and impact for Aotearoa and the world (Stats NZ, 2023). However, the full value and potential of this investment in research, science and innovation (RSI) has yet to be realised (REANNZ, NeSI, & NZGL, 2016). The Government is seeking to make the results of its investment more relevant and accessible to end-users (Saunders, 2022, p. 3), with a key lever being the programme of RSI reform known as Te Ara Paerangi Future Pathways (TAP) (MBIE, 2022a). The high-level goals of TAP are to realise “wellbeing for all current and future New Zealanders, a high-wage low emissions economy, and a thriving, protected environment through excellent and impactful research, science and innovation” (MBIE, 2022a, p. 9).

Data and datasets are central to this ambition. Data has been described as the world’s most valuable resource and, in Aotearoa, as a core national strategic asset (data.govt.nz, 2023). Public and private sector agencies hold a wealth of data about Aotearoa, its people, culture and environment. The Policy Framework for Government-held Information (PFGHI) (Booth, 2010) sets a policy-based foundation for the management of public sector information (PSI) but is no longer adequate for the 21st century RSI environment. With the release of the New Zealand Government Open Access and Licensing Framework (NZGOAL), a review of the PFGHI is being carried out to:

- support open and transparent government through active management and release of government’s non-personal information and data
- preserve the integrity and privacy of sensitive and personal information in government agencies, and
- create the conditions that encourage use and reuse of non-personal New Zealand government information and data for the benefit of the New Zealand economy and New Zealanders. (Booth, 2010; DPMC, 2023a)

Despite the plethora of activity, significant gaps remain. One of those is a system-wide view of the research data landscape – its scope, form, composition, opportunities and challenges. Some useful work has been undertaken on specific aspects of research data (Council of New Zealand University Libraries, 2016; Te Pūkai Tara, 2016a, 2016b, 2023), but a more complete picture is required to guide future strategic investment.

Internationally, considerable effort has already gone into mapping the research data landscapes in Canada (Alliance Research Data Management Working Group, 2022), United Kingdom (DARE UK, 2021), USA (Berman et al., 2016), Australia (Finkel, 2017; Switkowski, 2021) and Europe (Visionary Analytics et al., 2022). Collectively, they provide a useful context within which to situate an Aotearoa research data landscape review. There are, however, two distinct features of the Aotearoa context that set it apart and need to be accounted for. One is the central place of the Treaty of Waitangi | te Tiriti o Waitangi (te Tiriti), not only as Aotearoa’s founding document akin to a constitution, but as a central driver of government policy and practice, including those policies and practices that are part of the RSI system (MBIE, 2022a, 2023d). The other is the growing significance of Indigenous rights and responsibilities in relation to data in Aotearoa, often articulated as Māori data sovereignty (MDSov) (Te Mana Raraunga, 2018) and Māori data governance (MDGov) (Kukutai, Campbell-Kamariera, et al., 2023).

The purpose of this document is to provide a holistic review (the Review) of Aotearoa’s research data landscape. It covers four core areas:

- Te Tiriti o Waitangi and data sovereignty
- research data ecosystems
- research data infrastructure
- research data cultures.

Research data are data that are used as primary sources to support technical or scientific enquiry, research, or artistic activity; as evidence in the research process; and/or are commonly accepted in the research community as necessary to validate research findings and results. Research data may be experimental data, observational data, operational data, third-party data, public sector data, monitoring data, processed data or repurposed data (Alliance Research Data Management Working Group, 2022).¹ Much of the data used and generated by research in Aotearoa is not Findable, Accessible, Interoperable or Reusable; that is, it is not FAIR (Wilkinson et al., 2016b). In addition, much of the Māori data that is deployed or created through research is not meeting the requirements of MDSov (Te Mana Raraunga, 2018),² MDGov (Kukutai et al., 2023a), or even the minimum requirements of the international CARE Principles for Indigenous Data Governance (GIDA, 2019). The Review identifies these shortcomings and provides recommendations to remedy them so as to strengthen the system and advance shared aspirations for better outcomes.

METHODOLOGY

The methodology for this Review involved two broad steps. The first was a desktop review of the literature on research data and research data management (RDM) in Aotearoa and internationally. From this, key themes were identified and used to define a broad structure for mapping Aotearoa’s RDM landscape.

The second step involved engagement with the wider sector via Zoom, in-person conversations and targeted workshops.³ Select quotes from participants in those workshops have been incorporated into the various sections of the Review to highlight key themes and concerns. The following table lists the different foci of the targeted workshops.

Date	Workshop topic
28 March 2023	Kinds of data: Defining research data across its lifecycle Data contexts: Impact and value of research data

¹ For a fuller discussion of the definition of research data, see Appendix Three: Defining Research Data.

² Māori data refers broadly to digital or digitisable data, information or knowledge (including mātauranga Māori) that is about, from or connected to Māori. It includes data about population, place, culture and environment (Kukutai et al., 2023).

³ Initial engagement included virtual meetings with the Review core working team, a virtual meeting with the Committee on Data in Research (CoDIR) Forum, and a workshop at the 2023 eResearch Conference at The University of Waikato.

3 April 2023	Digital research infrastructure ecosystems and cultures
4 April 2023	Te Tiriti and its Indigenous implications – for individuals, groups and organisations
5 April 2023	People, skills, expertise and workforce Data culture conversations: Open and safe data in research
July 2023	Citizen science
July 2023	Crown Research Institutes (CRIs)

The conceptual approach of this Review is informed by mainstream concepts largely drawn from international literature. A tirohanga Māori (Māori perspective) of Aotearoa’s research data landscape would look quite different from the format used here. A Māori map of the RSI system is currently being undertaken by the MBIE-funded Māori RSI programme Kanapu (Kanapu, 2023) and will be available to inform future developments arising from the Review.

INTERNATIONAL REVIEWS

All of the other CANZUS countries (i.e., Canada, Australia and the United States) have undertaken reviews of their research data infrastructure, motivated by different concerns or agendas. This section provides a brief overview of the key themes/foci of those reviews along with those undertaken in the United Kingdom and Europe. Doing so helps to locate our Review within a broader international context, and to identify similarities and points of departure.

CANADA

The Current State of Research Data Management in Canada was published by the Alliance Research Data Management Working Group in 2022. It focused on the RDM ecosystem in Canada and the challenges and opportunities for supporting research excellence through RDM that is innovative, researcher-focused, inclusive and sustainable (Alliance Research Data Management Working Group, 2022). The review provided a ‘current state’ assessment of Canadian RDM with a specific focus on four areas: computing and storage infrastructure, interoperability, data service and governance. It covered a diverse range of entities including higher-education institutions, research organisations, research-funding agencies, scholarly publishers, academia-adjacent organisations, third-party service providers (commercial and non-profit) and international research organisations.

The report noted three distinct configurations of computing and storage infrastructure – active, repository and archival – to support the distinct stages of the data lifecycle. Infrastructure capacity depended on institutional capacity and there was significant variation in practices associated with data sharing across research domains. Repositories and institutions had begun pooling resources institutionally, regionally and nationally, often in the form of domain-specific repositories. The availability of archival storage able to accommodate long-term preservation of research data was a significant gap.

Interoperability is key to an effective RDM ecosystem. Interoperability requires common schemas, standards and protocols for collecting, organising and describing research data and supporting infrastructure. The report noted that operating frameworks are required that define the procedures, terms and relationships necessary to allow data to be exchanged unencumbered between digital

research infrastructures. Such frameworks allow new data, software and infrastructure to be developed and integrated by conforming to existing frameworks.

A wide range of data services are being developed to accommodate the growth in scale and complexity of services required, many being provided by commercial organisations. As these services scale, so will the infrastructure that is required to support the services to develop and grow. In line with this growth is the need for education and training in using these services and connecting infrastructure.

The final focus area of the report was on RDM governance, largely implemented by communities of practice who have developed guidance, policies and funding opportunities. It noted the need for consistent policies and requirements to enable researchers to adopt common practices and frameworks, both regionally and nationally.

The report noted that many of the RDM systems operate in isolation and require improved integration and adoption of shared standards, schemas and certifications for trusted interoperability. Continued consultation is required to further align and integrate organisations and services within Canada and also internationally.

UNITED KINGDOM

In the United Kingdom, the *Data Research Infrastructure Landscape: A Review of the UK Data Research Infrastructure* (DARE UK, 2021) was undertaken to support the development of a co-ordinated vision for digital research infrastructure in the UK, with a particular focus on managing sensitive data. The key findings were:

- UK-wide data and metadata standards for access and accreditation of data need to be developed and implemented.
- An issue even greater than data risk is gaining the trust of public data custodians and commercial organisations.
- Significant governance challenges are involved in co-ordinating across research infrastructures compared with the technical opportunities in developing federated data systems.
- Developing trust will require direct engagement with different groups (such as outreach activities), working in partnership with others in this space, and demonstrating examples of best practice interdisciplinary working.
- Retention of capability and capacity of those in the data space is required.
- Funding is an incentive that could be used to facilitate agreement about sets of standards and collaboration.

AUSTRALIA

The *National Research Infrastructure Roadmap* focused on the identification of “priority research infrastructure” required to underpin long-term, high-impact research in Australia (Finkel, 2017). The roadmap identified four layers of research infrastructure: institutional research infrastructure, national research infrastructure, landmark research infrastructure, and global research infrastructure (Finkel, 2017; Switkowski, 2021). The national, landmark and global research infrastructure comprised nationally significant assets, facilities and services to support leading-edge research and innovation; institutional

infrastructure, as the domain of individual institutions, was excluded (Finkel, 2017). The roadmap recommended the adoption of nine focus areas and priorities to strengthen Australia's economy, advance

societal benefit, improve their competitiveness, and build on their existing national capabilities: Digital Data and eResearch Platforms; Platforms for Humanities, Arts and Social Science; Characterisation; Advanced Fabrication and Manufacturing; Advanced Physics and Astronomy; Earth and Environmental Systems; Biosecurity; Complex Biology; and Therapeutic Development. It also recommended the establishment of a National Research Infrastructure.

EUROPE

European Research Data Landscape: Final Report was commissioned by the Directorate-General for Research and Innovation of the European Commission (Visionary Analytics et al., 2022). It had four focus areas. The first was on the use of data by researchers. The report found that the majority of researchers used under 10GB of data and up to 10 data sets, and most of their data is experimental and observational, with quantitative being more common than qualitative data. It also found that few researchers use data repositories but those who did so were more likely to understand and support the value of open science. The second focus area was researcher familiarity with the FAIR Principles (Wilkinson, Dumontier, et al., 2016). It found that researchers have a low familiarity and understanding of FAIR data, including in relation to data management plans (DMPs). The latter was greatly influenced by funding policies and their DMP requirements. The third focus was on data set FAIRness. The report found a high degree of variability in the application of FAIRness in repositories. The final focus area was the research data repository landscape. The report found that the vast number of respondents managed one data repository, which tended to be domain or discipline specific. Funding was not a major issue for repositories as they were supported by institutions or governments. Greater concerns also need to focus on digital and DMP skills of PhD students and early career researchers (ECRs), particularly in curricula relating to humanities and social science, to support them via data stewards with combined IT (information technology) competences.

The landscape report made four recommendations, to: (1) provide local support for research data management, (2) provide lifecycle support for data management planning and implementation, (3) facilitate the assessment of research data FAIRness and track progress towards FAIRenabling services and support, and (4) raise awareness of how FAIR benefits science and society.

UNITED STATES

The United States data landscape report *Realizing the Potential of Data Science: Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group* (Berman et al., 2016) made a number of important recommendations.

The first recommendation was in relation to a national data science research agenda. The report put the case for the creation of data science research centres and more investment in data infrastructure to further effective data sharing, use and lifecycle management. It identified the need for support for research into effective reproducibility and models that underlie evidence-based data policy, and expansion of funding for deep learning, smart environments and AI-powered and data-driven applications.

The second recommendation focused on a national data science education and training agenda. This would start with improving the design and development of a data science pedagogy and curricula and improving them within existing institutions. Data incubator programmes were needed, along with

public/private partnerships to engage with real-world problems and support PhD and postdoctoral fellowships in data science.

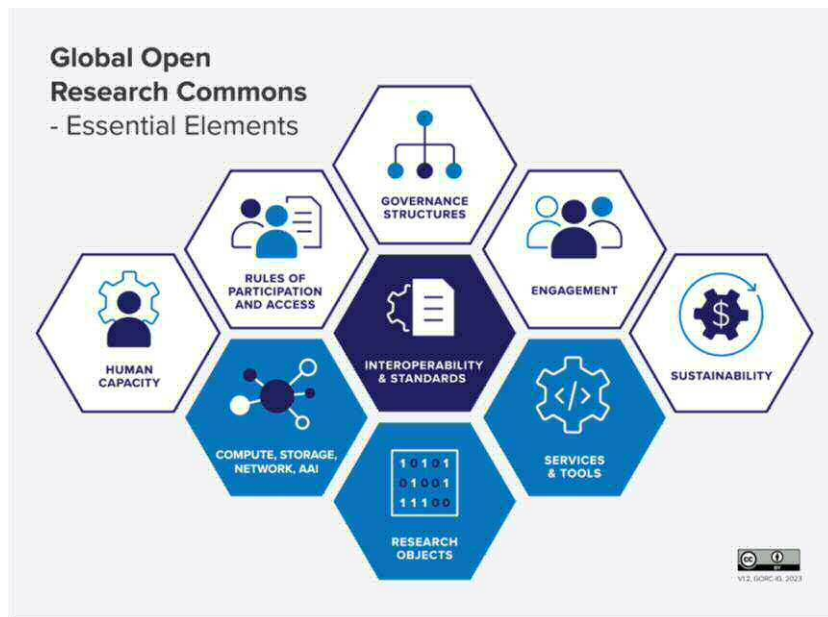
The third recommendation was on a national infrastructure to support data science and the acquisition of state-of-the-art and at-scale data sets for researchers. The infrastructure would support cutting-edge research and education, and work with libraries and domain repositories to develop innovative models of data stewardship. Best practice guidelines would be developed and evidence-driven investment and policy promoted.

The final recommendation was for new data-driven scenarios including strengthening the Internet of Things and the related data ecosystem. The report also noted the need to support research on radical hardware and software architecture that targets new and emerging data-intensive tasks.

GLOBAL OPEN RESEARCH COMMONS

Internationally, there is an increasingly rich and broad array of co-ordinated data infrastructures operating at various levels (country, continent, discipline, sector, mission). Often called “open science commons” or “data commons”, they provide a framework or platform to support an ecosystem of data, services, communities and outcomes (GORC IG, 2023).

Figure 1. Global Open Research Commons



Source: GORC IG (2023).

Prominent examples include the European Open Science Cloud, Australian Research Data Commons, Malaysian Open Science Platform, African Open Science Platform and the Nordic e-Infrastructure Collaboration, alongside an extensive range of more specific platforms in particular areas of science or scientific or national mission.

These communities are increasingly building their capabilities and performance through collaboration, such as through the Global Open Research Commons Research Data Alliance Interest Group (Research Data Alliance, 2021). Complexity and uniqueness challenge these communities’ ability to grow and sustain their performance and impact in isolation. There is enormous potential both to learn from the

substantial investments made globally and to explore the opportunities to respond to local needs with a willing community of colleagues.

AOTEAROA REVIEWS

Reviews of various aspects of research and data in Aotearoa have been undertaken but the emphasis has tended to be on research generally, rather than on the data landscape. Dietrich and Jones (2018) used value-chain mapping to understand the eResearch value chain of the eResearch community. This was workshopped and drawn together in the report *Understanding the eResearch Ecosystem in New Zealand*. It noted that the needs of researchers from different disciplines varied widely within Aotearoa, so any framework would need to capture requirements from multiple disciplines with a diverse range of needs (Dietrich & Jones, 2018).

The *National Research Data Programme: The Case for Research Data* was a future-oriented programme that brought together cross-institutional and cross-discipline views (REANNZ, NeSI, & NZGL, 2016). eResearch 2020 facilitated participants' foci on particular themes, research sector cloud strategies, skills gaps, institutional governance, research capabilities and the infrastructure needs of the National Science Challenges and the Centres of Research Excellence. Beyond this whole-of-system view, the following reviews have addressed various aspects of eResearch infrastructures or community-specific needs.

The *Kitmap – A Stocktake of Research Infrastructure in Aotearoa New Zealand's Government Research Organisations* found that infrastructure including laboratories and other types of infrastructures, such as field sites, physical and digital collections, computational resources, research vessels, and networks of monitoring sensors, reflected the fields of research prioritised by institutions (MBIE, 2022b). There was high demand for use of the infrastructure and the policies for infrastructure access varied significantly, including to external parties and often with fees attached. Most institutions funded their own infrastructure through grant overheads and commercial revenue funding, with a limited amount coming through direct government funding or funding from research grants.

Our Land and Water National Science Challenge: A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission used the concept of a *data ecosystem*, defined as “a system made up of people, practices, values, and technologies designed to support particular communities of practice” (Medyckyj-Scott et al., 2016). In such an ecosystem, data is valued as “an enduring and managed asset with known quality” (p. 21). The system required the infrastructure to deal with large data volumes, complexity and heterogeneity (often constrained by issues of privacy, intellectual property (IP) and licensing), and the use of a data management maturity model (DMMM). The DMMM would provide a shift from ad hoc approaches to managing and exchanging data (Level 1) towards the development and adoption of community-wide practices and standards for data sharing and data governance. Such changes would include: (1) changing the data management culture, (2) establishing a data analytical structure for the ecosystem (such as cloud-hosted), (3) increasing interoperability, and (4) building collaboration. The report noted that a lack of confidence and willingness to move towards significant transformation was often a barrier to change.

The *Research Data Management Framework Report* noted that although the concept of RDM has not changed, the environment in which research is conducted has, with researchers now generating extremely large volumes of data over very short periods of time (Wilkinson, Amos, et al., 2016). Data is now stored on volatile media in inaccessible locations and without any contextual semantics or clear lines of ownership, provenance or purpose. There is a significant risk that data will be lost, rendering the

publications, communications and discourse they generate undefensible and, in an academic context, useless. This signalled an important need for consistent RDM.

The *Infrastructure and Related Services to Support New Zealand's eResearch Future* report noted that eResearch is a prerequisite capability for a world-class RSI sector (MBIE, 2022c). Aotearoa has made modest investments in eResearch infrastructure but there are still capability shortages and limited uptake of eResearch, despite subsidisation of services. The data infrastructure needs to be accessible, responsive and flexible to the needs of all researchers, with transparent cost regimes in place to all users. Moreover, researchers should be supported by an eResearch community with skills and experience in all aspects of the eResearch lifecycle.

Te Ara Paerangi Future Pathways White Paper (TAP) proposes an ambitious agenda for reform of the Aotearoa RSI system (MBIE, 2022a). It sets out four high-level objectives for a future-proofed RSI system that includes: (1) enhancing research-policy linkages through long-term national research priorities (NRPs); (2) sustaining a talented, diverse and well-connected workforce; (3) having a system that is dynamic, high-impact and high-performing; and (4) embedding te Tiriti o Waitangi across all its parts. The reform has begun with the setting of NRPs by an independent panel (MBIE, 2023c) and the publication of a statement setting out MBIE's commitment to embedding te Tiriti across the RSI system (MBIE, 2023d). The NRPs are a Government tool to direct RSI resources to meet the most important challenges and opportunities for Aotearoa's social, environmental and economic wellbeing.

Finally, *The Future is Open: Establishing Wider Open Access for Research Publications in Aotearoa New Zealand* recommends the foundations for a long-term open-access (OA) strategy (Saunders, 2022). It argues that allowing research to remain locked behind paywalls deprives Aotearoa of the opportunity to realise maximum impact and value from its public investment. Limiting access slows the pace of scientific discovery, research commercialisation and the development of evidence-based public policy. It also serves to prioritise the interests of offshore publishing companies above the people who fund, conduct and contribute to research in the first place. Many of these recommendations have been put in place by the policy document *Kaupapahere Rangahau Tuwhera – Open Research Policy* (MBIE, 2023a), which is discussed in more detail in Section 3.

2. TE TIRITI O WAITANGI AND DATA SOVEREIGNTY

Having set out the rationale for this Review, and summarised international review efforts in the RDM space, this section focuses on the importance of te Tiriti and MDSov for the wider Aotearoa research data landscape.

TE TIRITI O WAITANGI

Signed in 1840 by hapū chiefs and representatives of Queen Victoria, te Tiriti o Waitangi is the foundational document of Aotearoa and regarded as part of its Constitution (Cooke, 1990). Te Tiriti establishes and guides the relationship between Māori and the Crown and a substantial body of Tiriti jurisprudence codifies some of the Crown’s specific duties to Māori. The Government recognises the importance of te Tiriti in establishing appropriate policy and procedure, and has provided detailed guidance to policymakers about how to take account of te Tiriti in policy development and implementation (Cabinet Office, 2019). Te Tiriti considerations must also be part of Government investment decisions.

One of the four key reform components set out by MBIE in the TAP white paper is embedding te Tiriti across the RSI system (MBIE, 2022a). The high-level statements on what such reforms envisage ‘embedding te Tiriti’ to mean include:

- promoting effective partnerships and suitable representation of Māori across RSI workforces, governance, leadership and management
- broad and purposeful investment in mātauranga and the promotion of a thriving ecosystem of Māori-led and community-led RSI activity, and
- recognising the rights, interests, duties and responsibilities of Māori in the allocation of resources for the realisation of RSI aspirations and adequate stewardship of rights.

This Review explicitly recognises and reflects the commitment to te Tiriti in TAP and in other keystone sector documents such as *Te Pūtahitanga: A Tiriti-led Science-Policy Approach for Aotearoa New Zealand* (Kukutai et al., 2021) and *A Guide to Vision Mātauranga: Lessons from Māori Voices in the New Zealand Sciences Sector* (Rauika Māngai, 2020). It also recognises the importance of MDSov (Te Mana Raraunga, 2018) and MDGov (Kukutai et al., 2023a) for all matters involving Māori data, including Māori research data. Some of these key concepts are defined below.

MĀTAURANGA

Mātauranga is a critical part of the research data landscape. The *Ko Aotearoa Tēnei* reports describe mātauranga Māori “as the unique Māori way of viewing themselves and the world, which encompasses (among other things) Māori traditional knowledge and culture” (Waitangi Tribunal, 2011, p. 6). Thus:

Mātauranga Māori incorporates language, whakapapa, technology, systems of law and social control, systems of property and value exchange, forms of expression, and much more. It

includes, for example, traditional technology relating to food cultivation, storage, hunting and gathering. It includes knowledge of the various uses of plants and wildlife for food, medicine, ritual, fibre, and building, and of the characteristics and properties of plants, such as habitats, growth cycles, and sensitivity to environmental change. It includes systems for controlling the

relationships between people and the environment. And it includes arts such as carving, weaving,

tā moko (facial and body tattooing), the many performance arts such as haka (ceremonial dance), waiata (song), whaikōrero (formal speechmaking), karanga (ceremonial calling or chanting), and various rituals and ceremonies such as tangihanga, tohi (baptism), and pure (rites of cleansing). (Waitangi Tribunal, 2011, p. 22)

While government agencies have a responsibility to ensure the active protection of mātauranga, they do not have the authority to control mātauranga or decide how it is used and shared (Mead et al., 2022). Mātauranga is held by Māori and belongs to Māori (Broughton & McBreen, 2015; Harmsworth & Awatere, 2013; Rauika Māngai, 2020).

This Review does not discuss the specific ways in which mātauranga is part of Aotearoa’s research data landscape or make recommendations pertaining to digital data containing mātauranga. Issues relating to the protection and use of mātauranga in research settings have already been well covered elsewhere (Mead et al., 2022; Rauika Māngai, 2020, 2022). Our view is that recommendations pertaining to the protection and use of mātauranga in the context of research are best left to mātauranga experts, kaitiaki (guardians) and knowledge holders.

MĀORI DATA IS A TAONGA

Māori data refers broadly to digital or digitisable data, information or knowledge that is about, from or connected to Māori people, language, culture, resources or environments (Kukutai et al., 2023a; Te Mana Raraunga, 2018). This reflects wider definitions of Indigenous Peoples’ data as comprising information and knowledge about Indigenous environments, lands, skies, resources and non-humans with which they have relations; information about Indigenous persons; and information and knowledge about Indigenous Peoples as collectives, including traditional and cultural information (Carroll et al., 2020).

Māori data is often described as a *taonga* – a tangible or intangible item or matter of special cultural significance to Māori (Riley, 2023; Ruckstuhl, 2023; Te Kāhui Raraunga, 2021; Te Mana Raraunga, 2017). Article 2 of te Tiriti guarantees the protection of iwi and hapū tino rangatiratanga over their ‘taonga katoa’. The question of whether something is a taonga is indicative of the strength of the Māori interest and therefore the standard of active protection required of the Crown (Waitangi Tribunal, 2021).

In its report into *Claims Concerning New Zealand Law and Policy Affecting Māori Culture and Identity* (WAI 262), the Tribunal defined taonga species as species over which whānau, hapū and/or iwi claim kaitiaki obligations through whakapapa and whose basis, history and content are set out in mātauranga (Waitangi Tribunal, 2011, p. 64). Developed over 40 generations, these relationships enabled practical uses including food production and harvesting (such as kūmara and tuna) and rongoā Māori (Waitangi Tribunal, 2011, p. 65). The Tribunal also defined taonga works and noted that each taonga work has a kaitiaki – those whose connection through whakapapa creates an obligation to safeguard the taonga itself and the mātauranga that underlies it. Examples of taonga works include mōteatea, karakia, carving, weaving, painting, constructions such as waka or wharehau and other crafts, and dramatic and musical works (Waitangi Tribunal, 2011, p. 30).

In its report on *The Comprehensive and Progressive Agreement for Trans-Pacific Partnership* (WAI 2522), the Tribunal did not specify which kinds of data are taonga in their own right but recognised that mātauranga included Māori rights and interests in the digital domain and this placed “a heightened duty on the Crown to actively protect those rights and interests, particularly in a field that is subject to rapid change and evolution” (Waitangi Tribunal, 2021, p. 50). It also recognised that “from a te ao Māori

perspective, the way that the digital domain is governed and regulated has important potential implications for the integrity of the Māori knowledge system, which is a taonga” (p. 53). While the Tribunal

did not state that all data is taonga, it is reasonable to expect that some kinds of data (including research data) will require more specific kinds of active protection given their sensitivity or value, and the contexts in which they are used. As a taonga, Māori data has its own mauri or life force (Hudson et al., 2018; Kukutai et al., 2023a) and requires the same respect as other living sources. The Crown’s responsibilities with regard to active protection of Māori data include influencing the broader settings within which the private sector collects, stores, uses and shares Māori data.

MĀORI DATA SOVEREIGNTY

Māori data sovereignty (MDSov) refers to the inherent rights and interests that Māori have in relation to the collection, ownership and application of Māori data (Te Mana Raraunga, 2018).

The MDSov network Te Mana Raraunga (TMR) notes that these rights and interests “derive from our inherent rights as Indigenous peoples, and unique relationships with land, water and the natural world. These rights are recognised in te Tiriti o Waitangi and the United Nations Declaration on the Rights of Indigenous Peoples” (Te Mana Raraunga, 2018, p. 1). MDSov can thus be seen as an extension of Māori sovereignty, rangatiratanga and mana motuhake (Kukutai & Cormack, 2019; Jackson, 2018; Ruckstuhl, 2023). In so far as MDSov emphasises Māori collective authority and agency, it extends beyond mainstream concepts of data sovereignty, which are primarily concerned with data residency and jurisdiction.

The MDSov Principles, and the concept of MDSov, have been widely referred to in research and policy-related contexts (see, for example, Bowen & Hinze, 2022; Greaves et al., 2023; Kukutai & Cormack, 2019; Morris, 2023; Oliver et al., 2022; Ruckstuhl, 2023; Sporle et al., 2020; Taiuru et al., 2023; Walter et al., 2021). Published by TMR, the 16 MDSov Principles are expressed in terms of six values:

1. Rangatiratanga | Authority (1.1 Control; 1.2 Jurisdiction; 1.3 Self-determination)
2. Whakapapa | Relationships (2.1 Context; 2.2 Data disaggregation; 2.3 Future use)
3. Whanaungatanga | Obligations (3.1 Balancing rights; 3.2 Accountabilities)
4. Kotahitanga | Collective benefit (4.1 Benefit; 4.2 Build capacity; 4.3 Connect)
5. Manaakitanga | Reciprocity (5.1 Respect; 5.2 Consent)
6. Kaitiakitanga | Guardianship (6.1 Guardianship; 6.2 Ethics; 6.3 Restrictions).

MDSov can only be exercised by Māori (including iwi, hapū) as the rights-holders through the retention and control of their data. Māori data governance is the mechanism that gives effect to MDSov, and is applicable to researchers, research institutions and research funders.

MĀORI DATA GOVERNANCE

Māori data governance (MDGov) refers to the principles, structures, accountability mechanisms, legal instruments and policies through which Māori exercise control over Māori data (Te Mana Raraunga, 2018). In 2023 the MDGov Model was published by Te Kāhui Raraunga, the operational arm of the National Iwi

Chairs Forum Data Iwi Leaders Group, as part of their Mana Orite relationship agreement with Stats NZ (Kukutai et al., 2023a). Designed specifically for use by the public service, the Model provides guidance for the system-wide governance of Māori data, consistent with the Government’s responsibilities under te Tiriti. The Model is intended to assist all agencies to undertake MDGov in a way that is values-led, centred on Māori needs and priorities, and informed by research.

The Model notes that a tremendous volume of Māori data has been collected by agencies over many decades, but with little transparency, poor access and few direct benefits to Māori. Māori authority over government-held Māori data is generally weak to non-existent. The Model defines eight critical areas of data governance in the form of Data Pou or pillars: data capacities and workforce development; data infrastructure; data collection; data protection; data access, sharing and repatriation; data use and reuse; data quality and system integrity; and data classification.

The Model does not cover every element of data governance. Instead it focuses on key priorities and actions, against which agencies can assess their level of data maturity for each Data Pou. The Model explicitly recognises the need for changes to system leadership, policies and legal settings so that Māori can exercise authority over Māori data to reduce unethical data use and strengthen outcomes for individuals, whānau and communities. The Model also identifies the need for strategic investment in a ‘mana motuhake’ data system that sits outside the public sector to ensure iwi and hapū sovereignty over iwi and hapū data. Although designed for the public service, the Model lends itself to the public sector more broadly, including universities and CRIs.

INDIGENOUS DATA SOVEREIGNTY AND DATA GOVERNANCE

Beyond Aotearoa, the Indigenous data sovereignty (IDSov) movement has begun to have a significant impact on research, policy and Indigenous advocacy. Much like MDSov, IDSov refers to the right of Indigenous peoples and tribes to govern the collection, ownership and application of their own data (Caroll, Rodriguez-Lonebear & Martinez, 2019). Since the publication of the first book on the topic in 2016 (Kukutai & Taylor, 2016), IDSov scholarship has grown exponentially, covering a diverse range of fields from the law (Tsosie, 2019), information privacy (Kukutai et al., 2023b) and trade (Mika et al., 2023) to AI (Dobson & Fernandez, 2023), genomics (Hudson et al., 2020) and research (Garba et al., 2023; Jennings et al., 2023).

In recent years there has been increasing recognition of the importance of IDSov (FNIGC, 2020).⁴ The UN Special Rapporteur on the right to privacy has highlighted the importance of IDSov in two reports – one relating to Big and open data, and the other on the use of personal health data (Cannataci, 2018, 2020). In the 2018 report, he noted that data “is a cultural, strategic and economic resource for indigenous peoples” and that “existing data and data infrastructure fail to recognize or privilege indigenous knowledge and worldviews and do not meet indigenous peoples’ current and future data needs” (p. 13).

⁴ See, for example, the First Nations Data Governance Strategy (FNIGC, 2020) and the First Nations Principles of OCAP®: <https://fnigc.ca/ocap-training/>

He also called on national governments and private corporations to recognise “the inherent sovereignty of indigenous peoples over data about them or collected from them” (p. 9).

The Canadian *United Nations Declaration on the Rights of Indigenous Peoples Act Action Plan* has an entire section dedicated to IDsov (Department of Justice Canada, 2023). Section 30 states that, among other things, the Government of Canada will:

... continue to support Indigenous Data Sovereignty and Indigenous-led data strategies through legislative, regulatory and policy options to help ensure that First Nations, Inuit, and Métis have the sufficient, sustainable data capacity they need to control, manage, protect, and use their data to deliver effective services to their peoples, tell their own stories, participate in federal decision-making processes on matters that impact them, and realise their respective visions for self-determination.

The Global Indigenous Data Alliance (GIDA), which represents the collective voice of IDsov networks in the CANZUS states and Scandinavia, published *Indigenous Peoples’ Rights in Data*, which describes the specific rights that support Indigenous Peoples’ aspirations for control of data and self-determined research activities (GIDA, 2023).

GIDA are also the stewards of the well-known CARE Principles for Indigenous data governance (RDA IG, 2019). CARE stands for **C**ollective Benefit, **A**uthority to Control, **R**esponsibility and **E**thics. Developed in 2018, the CARE Principles explicitly recognise the tensions between the push for open data and Indigenous Peoples’ assertion of greater control over the application and use of Indigenous data and Indigenous knowledge for collective benefit (Carroll et al., 2020, p. 3). The Principles empower Indigenous Peoples by shifting the focus from regulated consultation to value-based relationships that position data approaches within Indigenous cultures and knowledge systems to the benefit of Indigenous Peoples (Carroll et al., 2020). The CARE Principles are complementary (rather than antagonistic) to the FAIR Principles (see Section 3). While assessing FAIRness is an exercise that an individual researcher or team can do, the CARE Principles require engagement with people to address the cultural, ethical, legal and social dimensions associated with the intended uses of the data set (Carroll et al., 2020, 2021). The CARE Principles have been either endorsed or adopted by the Research Data Alliance, the UNESCO Recommendation on Open Science, IEEE Recommended Practice for the Provenance of Indigenous Peoples’ Data, and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) Code of Ethics for Aboriginal and Torres Strait Islander Research.

GIDA recently published the *Indigenous Data Governance and Universities Communiqué* which called on universities to take a number of actions including implementing DMPs and data management strategies that have IDGov principles and mechanisms embedded throughout, and allocating adequate resources for Indigenous Peoples to govern their data on their own terms (Prehn et al., 2023). Much of the data that is being generated from research about or involving Indigenous Peoples and their territories is not discoverable or accessible, and Indigenous research governance of Indigenous data has been limited. Garba et al. (2023) provide guidelines on how researchers and institutions can recognise and uphold research sovereignty, and how communities can set out their expectations with regard to recognising sovereign relationships and upholding their rights and interests in data, including sovereignty of research data (Garba et al., 2023).

WORKING WITH LOCAL AND INTERNATIONAL MODELS OF IDSOV AND IDGOV

The CARE Principles should be seen as a starting point rather than the end point for data governance in Aotearoa. When it comes to Māori data, the MDSov Principles and MDGov Model set a significantly higher threshold of authority, protection and benefit, being grounded in te Tiriti (empowering both rangatiratanga and partnership) and tikanga Māori values.

In the context of international RSI collaboration, the CARE Principles have the advantage of being relatively well socialised across different national contexts. They provide a common framework for researchers, communities and funders so they can agree a shared approach to governing Indigenous data in contexts that may be quite different (e.g., Horizon Europe).

Given the significance of MDSov and MDGov in Aotearoa, and of IDSoV and IDGov internationally, key actors in our research data landscape need to take meaningful steps to give effect to MDSov and MDGov in terms of their policies, practices and processes. These actors include the Ministry of Business, Innovation and Employment (MBIE) as the major RSI policy agency and funder, as well as research institutions including universities, CRIs and independent research organisations. Implementing MDGov over Māori research data is not just about meeting te Tiriti accountabilities; it is also about shifting the research data ecosystem to be more responsive, ethical, transparent and effective.

3. RESEARCH DATA ECOSYSTEMS

DATA ECOSYSTEMS

An important starting point in evaluating the research data landscape is to understand the data ecosystems and data lifecycles that occur within the landscape (SGIG DSC, 2023). There are some general features of data ecosystems and lifecycles that are universal, as well as features that are unique to Aotearoa.

In general, data ecosystems are networks or sets of networks where autonomous actors directly or indirectly consume, produce or provide data and other related resources (Oliveira & Lóscio, 2018). More specifically, the term *data ecosystem* encompasses:

- policies regarding data management planning, data custodianship and curation, legal frameworks, and the use of externally sourced data
- procedures and processes to execute those policies and manage data
- a data governance framework and organisational structures
- engagement with data consumers and stakeholders, and
- technology platforms that will support data collection, storage, description, analysis, linking, delivery and curation. (Medyckyj-Scott et al., 2016, p. 5)

Data ecosystems exist at various levels including community and regional, national and global. While separate, there may be interaction between these systems (Medyckyj-Scott et al., 2016, p. 5). An ideal data ecosystem has been described as one that is “mature”:

A mature data ecosystem is a distributed, adaptive, open socio-technical system with properties of self-organisation, scalability, and sustainability that turns data into information and knowledge. It comprises a system made up of people, practices, values, and technologies designed to support particular communities of practice. (Medyckyj-Scott et al., 2016, p. 5)

Within Aotearoa’s emergent research data ecosystem, networks and groups of actors operate sometimes in isolation and at other times in collaboration with others (Workshop 3). There is no current national uniformity or general guidance on data collection, management, curation, sharing, use and reuse. Nor is there national-level leadership or investment in a data ecosystem.

The literature and the feedback from our workshops suggest that the ecosystem in Aotearoa is growing and dynamic but is still in its early stages compared with the other CANZUS countries. Core data needs relating to data storage, cataloguing (such as through metadata), and sharing are not being adequately met. For the majority of research, the national management of data is maintained in an ad hoc manner (Workshop 1). There are some prominent exceptions, one being the Nationally Significant Collections and Databases funded through the Strategic Science Investment Fund Infrastructure appropriation.⁵ There is

⁵ <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/funding-information-and-opportunities/investment-funds/strategic-science-investment-fund/funded-infrastructure/nationally-significant-collections-and->

also the work carried out by the national statistics office, Stats NZ, through data.govt.nz and the Integrated Data Infrastructure (IDI), although it lacks a high profile in the RSI sector.

RESEARCH DATA MANAGEMENT

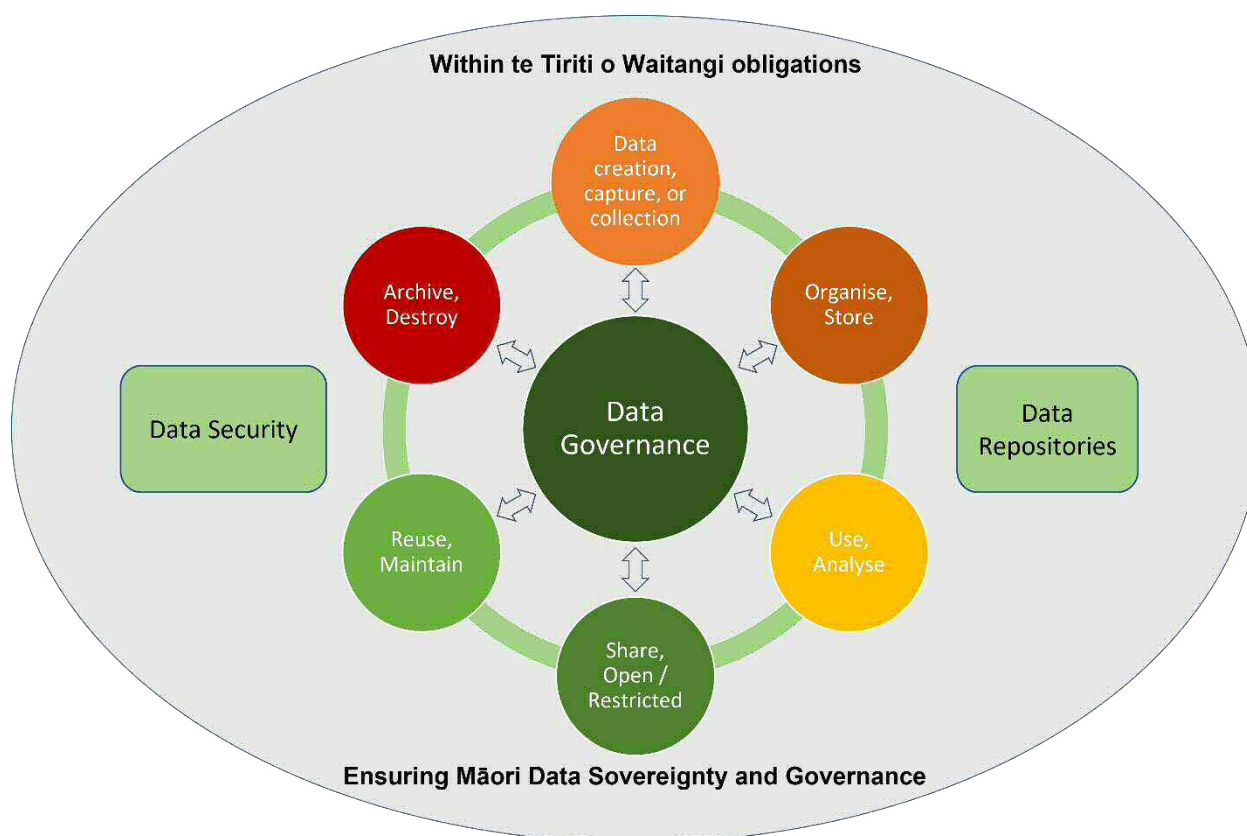
Research data management (RDM) refers to the processes of planning and undertaking the collection, organisation, management, storage, backup, preservation and sharing of data, before, during and after a research project. Research has always required some form of data management but the environment and technology in which research is conducted has become far more complex (Wilkinson, Amos, et al., 2016).

Key to RDM is the concept of the *data lifecycle* – the different stages a unit of data undergoes from initial collection through to when it is no longer considered useful and deleted. RDM is an ongoing process of collecting, using, storing, transferring and destroying or permanently archiving data, which is continually repeated with different data sets (Data.NSW, n.d.). The understanding of the data lifecycle varies across fields and contexts but there are some commonalities. A core part of the data lifecycle is having effective data lifecycle management (DLM). *Data lifecycle management* concerns the best practices throughout the various stages of the lifecycle: production, data cleansing, data management, data protection and data governance (Kamaly, 2022). Participants in one of our workshops provided a more nuanced conceptualisation of the data lifecycle as comprising:

1. Ideation / Planning / Design
2. Create / Gather / Obtain
3. Analyse / Interpret / Visualise / Understand
4. Disseminate / Publish / Report
5. Enable discovery and reuse by others. (Workshop 1)

While the data lifecycle tends to be conceptualised as a sequential process, we think it more productive to see the data lifecycle in Aotearoa as a dynamic system of data governance considerations and decisions (see Figure 2).

Figure 2: Aotearoa data lifecycle



Source: NZCoDIR.

In this model, we aim to highlight that data governance is a continuously interactive process that operates throughout the entire data lifecycle. Consequently, effective RDM planning should integrate strong decision-making mechanisms at every stage of the data lifecycle, with well-defined and mutually agreed-upon responsibilities for each step. Within this dynamic model, there exists an opportunity to incorporate both MDSov and MDGov. This can be achieved by leveraging the substantial resources and guidance already available and employing various mechanisms, such as data management plans (DMPs) (Workshop 1).

The general feedback from our workshops is that RDM in Aotearoa is inconsistent and ad hoc and lacks a clear, co-ordinated approach. Universities New Zealand identified the need for a long-term strategy to design a robust RDM process that can be integrated into current practice (Wilkinson, Amos, et al., 2016). Most universities have begun to encourage researchers to produce a DMP as part of their proposal and project, providing tools and assistance to support them to do so (Library Research Services, 2023; University of Auckland, 2023c; University of Otago Library, 2016). However, in most instances, a DMP does not appear to be an explicit requirement that is monitored for compliance – although there are some exceptions. The University of Auckland’s new Research Data Management Policy sets out the responsibilities of the university, researchers and supervisors for the management of research data. It requires a DMP for “sensitive or restricted data or where this is required by the University, a funder, ethics approval processes, research data provider or other external party” (University of Auckland, 2023c). The policy notes that research data should be managed in ways that are consistent with:

- international standards for FAIR data and open research that are increasingly required by funders, data providers and publishers
- the University’s obligations under Te Tiriti o Waitangi and commitment to becoming a Māori data sovereignty organisation, and
- the CARE principles for the governance of Indigenous data, including Pasifika data, and legal, ethical and protective security requirements for research data.” (University of Auckland, 2023c)

Though government funding agencies have not yet required researchers to develop DMPs as a condition of funding, it is now strongly recommended by MBIE (MBIE, 2023a). Many journals also require an open data set as a condition of publication.

The *Data Maturity Model* (DMM) sits alongside RDM as a framework for evaluating the maturity level of an organisation’s data practice capabilities. At present there is no national DMM and its use is limited across institutions. A potential first step is to document what already exists and the barriers to linking them into a national model (Workshop 1). Some research institutions have DMM models; for example, the University of Auckland has a RDM capability maturity model (Armstrong et al., 2021).

There is a need to develop RDM processes that are suitable for the Aotearoa data ecosystem and that account for MDSov and MDGov. Like ethics processes, RDM processes are approved at the time of project approval. For the DMP to be effective, it needs to be an active process throughout the life of the project and beyond into a post-project system so that the data is stored and curated appropriately.

MBIE has previously made statements of intent about RDM and data governance but to date there has been little tangible action in the form of implementation, which is in contrast to Australia (Workshop 1). There are some promising signs of change, including the programme of reform set out in TAP and *Kaupapahere Rangahau Tuwhera – Open Research Policy* (MBIE, 2023a).

DATA DISCOVERY

It is not known how much research data currently exists in Aotearoa as no systemic stocktake has been conducted. There have been some attempts, but they became too complex and were never finished (CRI workshop). There is no catalogue system of projects or data/data sets. Many researchers have stored their data on either their personal or institutional systems with limited tagging of what has been recorded and where. There are no uniform procedures in place in Aotearoa to capture how data is used, what data is used, and the ethics in doing so. As one workshop participant noted:

Ultimately, it is not fully clear what data there is as much of it is with a researcher or institutional location and not accessible or documented. There is a need for a national research data maturity assessment. (participant, Workshop 1)

The situation is even more parlous for Māori. Much of the data generated from research about or involving Māori individuals, collectives and their environs is not readily discoverable or accessible to those external to the research team, making it extremely difficult for Māori to discover or access Māori data, let alone exercise any meaningful decision-making authority over it. While the use of Māori metadata (see the next section) is necessary and desirable, institutions first need to take some level of responsibility for identifying where Māori data is held within their institution and then implementing MDGov to ensure that such data is being governed and managed in ways that are culturally appropriate. Aotearoa is not alone in this regard. The recent GIDA communiqué calls on all universities to “identify Indigenous Peoples’ data currently held

by the institution or its research partners” and to “introduce and strengthen University policies to align with IDSoc and IDGov and ensure Indigenous Peoples’ governance of, access to, and future use of their data now and into the future” (Prehn et al., 2023).

Beyond Aotearoa there are some international benchmarks or exemplars of data discovery, mostly involving a single or field-dependent data portal. Canada, through the Digital Research Alliance of Canada, has introduced a system called Lunari.⁶ It is a scalable, bilingual national research data discovery service that provides a single point of search for Canada’s multidisciplinary research data held in a variety of repositories, including those of post-secondary institutions, departments at all levels of government, research organisations and national repository initiatives. Information about data sets in the form of metadata records are harvested from repositories and made available for discovery. There are over 80,000 data sets from over 100 Canadian repositories and data collections currently indexed by Lunaris. Additionally, the Lunaris index is harvested by other discovery aggregators, including OpenAIRE Research Graph, Data Citation Index and Proquest Central Discovery Index.

Another international system is the POLDER Federated Search (World Data System initiative).⁷ This is an example of an interdisciplinary research focus on a particular region, in this case the polar region. It is a federated metadata search system for the polar regions that simplifies the discovery of polar scientists. It is a collaboration between the Southern Ocean Observing System, Arctic Data Committee, and Standing Committee on Antarctic Data Management.

In Australia, the Bioplatforms Australia Framework Initiatives are national collaborative projects that use integrated omics infrastructure to generate high-impact data and knowledge resources to support some of Australia’s biggest scientific challenges.⁸ The resulting data sets are stored within (or linked to) its data portal and are all accompanied by a rich set of metadata to enhance reuse.

There are international systems established to bring together open data. An example of such a platform is Dryad,⁹ an open data publishing platform and a community committed to the open availability and routine reuse of all research data to benefit society worldwide. It is a signatory to the Principles of Open Scholarly Infrastructure (POSI).¹⁰

A further example is the Dutch OpenAIRE Connect.¹¹ The portal covers publications, data sets and projects of research institutes in the Netherlands to connect researchers to a broad audience of national and international researchers, journalists, policymakers and companies. It gathers research output (publications, data sets, software and other research products) from a range of institutional repositories, national and institutional research information portals, data repositories and software repositories.¹²

There are a number of data systems and data sets in Aotearoa. Some of the data sets include weather data (MBIE, 2018), Destination Earth, which focuses on developing a single global infrastructure with a range of digital twins specific to particular use cases (European Centre for Medium-Range Weather Forecasts, n.d.), and the Integrated Data Infrastructure (IDI) provided by Stats NZ . There are also larger collections of data in larger organisations and government agencies (Workshop 1). If easily identified, the

⁶ <https://alliancecan.ca/en/services/research-data-management/lunaris>

⁷ <https://search.polder.info/about/>

⁸ <https://data.bioplatforms.com/about>

⁹ <https://datadryad.org/stash/about>

¹⁰ <https://openscholarlyinfrastructure.org/>

¹¹ <https://netherlands.openaire.eu>

¹² <https://www.uksg.org/newsletter/uksg-enews-543/new-research-portal-netherlands-openaire>

created data will be able to be reused by researchers (Workshop 1). Many of the existing research data sets do not currently provide for MDSov or MDGov.

DATA ON RESEARCH PROJECTS

Thousands of research projects are carried out in Aotearoa every year, funded in a variety of ways. Those that have been publicly funded should be readily discoverable but too often are not.

There are some examples where organisations or networks have collated information on research projects in particular domains or contexts. One example is the IDI stewarded by Stats NZ.¹³ The IDI is a large research database that holds de-identified microdata about people and households. Stats NZ states that “researchers use the IDI for research in the public interest to improve outcomes for New Zealanders” (Stats NZ, 2022). More than 1000 projects have used Stats NZ microdata, including numerous projects that have focused on Māori outcomes (Stats NZ, 2022). The projects and associated publications from the IDI as well as the Longitudinal Business Database (LBD), other data sets available in the Data Lab, and Confidentialised Unit Record Files (CURFs) are listed on a searchable Stats NZ website (Stats NZ, n.d.).

Another example of collated information about research projects is the stocktake of cardiac research. The stocktake revealed a registry of a ‘library’ or directory of the projects that have already been run and the data associated with them and who to contact about them (Workshop 1).

There are several benefits of centralising information about research projects. One obvious one is that it reduces unnecessary duplication (unless necessary for scientific rigour). Māori, in particular, have been over-researched, with the same information being extracted time and again (Smith, 1999). Another benefit is that centralised information provides those outside of research institutions with clear information on what research has been undertaken on areas of interest, or on topics affecting their communities. One potential benefit of implementing a centralised approach would be to better understand whether projects have engaged with or produced mātauranga Māori or Māori data. This information could be useful to Māori collectives for a wide range of purposes and help make visible data and information that requires active governance and protection.

Many research proposals are written annually (many more than are funded), and contain important background information and ideas. Researchers are hesitant to share their proposal ideas with those other than their collaborators, given the competitive nature of funding which encourages people to hold ideas (as a form of IP) close to their chest. Sharing research proposals in an open or semi-open way would require a high degree of trust and sensitivity.

The New Zealand Research Information System (NZRIS) has the potential to provide a more comprehensive view of the research being undertaken in Aotearoa. Stewarded by MBIE, NZRIS is a metadata repository about research. It will be able to show what projects have been funded, their foci and subject areas, the funding mechanism and amount, the researchers involved, the community partners, project outputs/publications, and more (MBIE, 2019). NZRIS will begin with data provided by MBIE, the Health Research Council of New Zealand and the Royal Society of New Zealand Te Apārangi. Initially NZRIS will only include funding data from the funding organisation, but will eventually also include the data about the research activity as well.

¹³ Stats NZ is legally mandated to collect, store and operationalise administrative data on behalf of the Government and its agencies. A growing number of administrative data sets are linked and used within the IDI.

METADATA

Metadata is data about data. *Metadata* provides the infrastructure to represent the descriptive, structural, administrative, reference, statistical and legal elements of data. It enriches the data with information that makes it easier to find, understand, authenticate, trust, use and manage information and records (Archives New Zealand, 2023). An important role of metadata is to describe data provenance.

Metadata is a critical component of Aotearoa’s research data infrastructure but, like other aspects, is somewhat ad hoc and inconsistently applied. Discussions and proposals are needed to understand how metadata scaffolding can and could be done and what sorts of systems are needed to support it. Minimum standards for metadata (Archives New Zealand, 2023) do not currently specify a field for Indigenous or Māori knowledge – it is entirely voluntary. That leaves open the loss of any Māori data or awareness that there could be relevance for Māori. It is critical that Māori data reflects Māori and is captured in a way that is relevant to Māori (Science for Technological Innovation National Science Challenge, Data Iwi Leaders Group, & Victoria University of Wellington, 2018). As a minimum, there ought to be specific fields for Māori data that include provenance, involvement and consent.

Having good metadata is also important for realising the value of data.

Value might be looked at as a bit like currency – over time put some money in the bank and compounding interest grows its value. If the data is all in varying currencies (i.e., not common standards/metadata), then each little bit of data has some value, BUT if the data was all in the same ‘currency’, then it can be aggregated and the value grows even more. (Workshop 1)

DOIS

A *digital object identifier* (DOI) is a digital identifier of an object – any object: physical, digital or abstract – that enables reliability in identifying and accessing that particular item. DOIs create a permanent reference, making research outputs or data in FAIR, and enable citation of research outputs or data, and tracking of impact metrics. A DOI can distinguish between objects with similar titles or different versions (University of Auckland, 2023b). The National Library of New Zealand holds the hub for New Zealand DOIs (NZ DOI). Membership is open to any New Zealand organisation as long as they meet the criteria (National Library, 2023).

DOIs do not set out or provide for any Māori rights or interests that may be connected with the research. Tools such as Local Contexts’ Notices and Labels are a method to overcome these issues (see below).

PIDS FOR INSTITUTIONS/PROJECTS

As well as *persistent identifiers* (PIDs) such as ORCID (Open Researcher and Contributor ID), there are other PIDs for organisations. The Research Organization Registry (ROR) is an example of a global, community-led registry of open PIDs for research organisations (ROR, n.d.). RAiD is an Australian example of a project-based PID (RAiD, 2022). It provides persistent, unique and resolvable identifiers for research projects based on the global handle system. Its metadata envelope includes persistent identifiers for funders and grants, organisations and institutions, collaborators and contributors, articles and data, and tools and services. (See also, the discussion of PIDs in Section 5.)

METADATA CHALLENGES FOR MĀORI

The lack of culturally informative metadata is a major barrier to the implementation of MDSov and MDGov, but emerging initiatives offer solutions. As the MDGov report notes, “Data that is classified as Māori data ... and that is not subject to restrictions, should have appropriate metadata that increases its findability” (Kukutai, Campbell-Kamariera, et al., 2023, p. 38).

Archives New Zealand, which is the regulator of information created by the public sector, has entered into formal agreements with iwi to better understand what information and taonga tuku iho the Archives hold that is connected to iwi and hapū and to Māori more broadly (Archives New Zealand, 2018).

Requiring Māori metadata will improve access to such information and enable iwi and hapū to search and locate their data that has collected by other institutions. Māori metadata is more intuitively aligned with a tirohanga Māori and can identify specific iwi, hapū, marae, whenua, people, whakapapa, koawa, te reo Māori, mita and place names. The hope is that implementing the use of Māori metadata to identify taonga tuku iho will improve access to, and use of, information and knowledge relating to iwi, hapū and other Māori collectives.

Ensuring Māori metadata is recorded and embedded is critical to make explicit the kaitiaki of the data and their rights and interests. This makes it easier for kaitiaki to influence decisions about access and sharing, and to realise any benefits that may be generated from the data.

LOCAL CONTEXTS

Local Contexts is a global initiative that supports Indigenous communities with tools that can reassert cultural authority in heritage collections and data (Local Contexts, 2023). The Local Contexts hub is a web portal that enables the customisation of labels by Indigenous communities and the generation of notices by institutions and researchers.

Notices are tools for institutions, repositories and researchers to support ethical use and reinforce relationships with Indigenous communities. They also work to educate the public around Indigenous rights and interests in historical and future collections and data.

Traditional Knowledge (TK) and Biocultural (BC) labels establish Indigenous cultural authority and governance over Indigenous data and collections by adding provenance information and contextual metadata (including community names), protocols and permissions for access, use and circulation of the data.

Labels and notices can be applied to Local Contexts projects created within the hub, depending on the account type. By focusing on Indigenous IP and IDSov, Local Contexts helps Indigenous communities repatriate knowledge and gain control over how data is collected, managed, displayed, accessed and used in the future.

In Aotearoa, TK labels have been added to all the data in the Systematics Collections held by Manaaki Whenua | LandCare Research. Some iwi, including Whakatohea, have attached BC labels to the records of data that are connected to them (Manaaki Whenua, 2023).

The Institute of Electrical and Electronics Engineers (IEEE) – the leading developer of industry standards in a broad range of technologies – is currently supporting the development of recommended practice to establish and define a common set of parameters by which the provenance of Indigenous Peoples’ data should be described and recorded (IEEE, 2020).

OPEN DATA AND OPEN RESEARCH

OPEN DATA

Open data is data that is openly accessible, exploitable, editable and shared by anyone for any purpose. Open data is licensed under an open licence. The open data movement had its genesis in the early days of the internet when people started realising the potential of sharing information online. Prominent advocates for open data include the Open Data Charter – a collaboration between over 170 governments and organisations working to open up data based on a shared set of principles – and the Open Data Institute.¹⁴

Our government has promoted an open data approach for more than a decade and signed the Open Data Charter in 2017. Along with Canada, Australia, the United States and Nordic states, Aotearoa is a member of the Open Government Partnership founded in 2011 to promote transparent, participatory, inclusive, accountable governance. The concept of open data deployed by the government is:

... data that anyone can access, share and use. Data must be open, trusted, authoritative, well managed and readily available. Opening up data for reuse has widespread benefits to government, the private sector and the public. (Digital.govt.nz., 2019)

The New Zealand Data and Information Management Principles are a set of seven principles under which the government releases its open data. None of the principles recognise (or have been amended to recognise) Māori rights and interests.

OPEN RESEARCH

Open research (also referred to as open science and open scholarship) describes a range of processes and practices that aim to improve the reach, reproducibility and rigour of research. These processes and practices differ according to discipline and can fall under the broad areas of open access, research data management and open data, among others. Overseas, open research position statements are commonplace; however, the University of Waikato is the only university in Aotearoa to have such a statement (University of Waikato, 2023). The FAIR and CARE Principles should be an integral part of the process of open research, with more value and effort put into that domain.

The emerging consensus across the sector is that open research is critical to advancing future research and innovation in Aotearoa. This has been highlighted most recently through New Zealand's Association to Horizon Europe, which introduces a new open-science policy requiring mandatory open access to publications and the application of open-science principles throughout the programme (Directorate-General for Research and Innovation (European Commission), 2021). It is important that researchers understand how to make data open while protecting particular elements as required. For example, application of the CARE Principles within the *te Tiriti* context has been recognised in the articles of Association with Horizon Europe.

Globally, there have been moves to encourage, or in some cases obligate, publicly funded research to have their research outcomes, including data, available through open access. Though many researchers welcome this, barriers remain to achieving such a goal.

Plan S, a global initiative aimed at accelerating the transition to open-access publishing, has garnered support from numerous research funders worldwide. These funders have made significant commitments

¹⁴ <https://theodi.org/>

to reshape the scholarly publishing landscape. By signing up to Plan S, these organisations commit to the principle that research results should be openly accessible to the public. Obligations for grantees typically include publishing their research outputs in compliant open-access journals or platforms. This entails making research articles and data freely available to all, without restrictions, enabling broader dissemination and maximising the societal impact of research.

MBIE's Open Research Policy mandates all MBIE (and related government-funded) projects to have their publications deposited in open-access portals (MBIE, 2023a). The policy also strongly recommends that research data arising from funded projects be made openly available, but notes some exceptions including IDSoV considerations and copyright restrictions. Researchers are also encouraged to produce a DMP that includes measures to safeguard and protect Indigenous or other sensitive data.

FAIR AND CARE PRINCIPLES

The FAIR Principles for scientific data management and stewardship were published in 2016. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of digital assets. Through the Principles, FAIRness ought to be applied by both human-driven and machine-driven activities, and is a critical component to consider in the data management and stewardship process – from researchers and data producers to data repository hosts (Wilkinson, Dumontier, et al., 2016). FAIR focuses primarily on the data itself rather than the conduct of the producers, to enable common visions of all areas of the data publishing ecosystem (Wilkinson, Dumontier, et al., 2016).

The FAIR Principles are designed to provide for rigorous management and stewardship of digital resources, to the benefit of the academic community (Wilkinson, Dumontier, et al., 2016). A FAIR Data Maturity Model has been created to harmonise FAIR assessments (Bahim et al., 2020).

Though many organisations are now discussing the FAIR Principles in relation to data, it is not clear the extent to which they are being applied within organisations. As a participant noted in Workshop 1: “I wonder what we would see if we did an analysis of the FAIRness of data.” Although the FAIR Principles have been out for some time, there has been limited socialisation of them in Aotearoa and how they apply to researchers.

The CARE Principles for Indigenous Data Governance, already described in Section 2 of this Review, sit alongside the FAIR Principles as a mechanism to ensure appropriate governance over Indigenous data in the context of open data and open research.

4. DATA INFRASTRUCTURE AND SYSTEMS

As the growth of the data industry expands, so does the infrastructure that enables it to function and flow nationally and internationally. The government has invested in some of this infrastructure, but more could be done. A government consultation is currently underway aimed at enhancing the resilience of Aotearoa critical infrastructure (DPMC, 2023b).

DATA INFRASTRUCTURE

Data infrastructure is often defined widely to include instruments, archives/storage, knowledge-related facilities, computational systems, communication networks, software, security and metadata (Workshop 2). Such infrastructure enables “the possibility to describe, create, store, share and reuse data/artefacts. Infrastructure needs to be sustainable and easy to access” (Workshop 2).

Data infrastructure can occur at various levels, from individual, institutional and consortia to regional and national. Having an infrastructure in place enables knowledge of what data is available and where it is. A robust infrastructure requires policies and procedures to ensure ethical, transparent and trustworthy access and use of such data. Using infrastructure/tools enables *interoperability*; that is, allows data to be transferred between infrastructure (e.g., when new infrastructure is available or when existing infrastructure is discontinued).

Some of the important characteristics to consider include whether the infrastructure is:

- fragmented
- tailored/bespoke
- interoperable
- hierarchical (research group, research organisation, national, international)
- highly collaborative
- highly dynamic
- stable, upgradeable
- easy to access, and
- sustainable.

Data infrastructure also comprises nonphysical and digital elements, including people, capabilities, physical, virtual, constructed, active, maintained, cybersecure, service levels, value propositions, business models, operating models, governance, cultures, collaboration and incentives.

There has not been a substantive review or stocktake done of the data research infrastructure in Aotearoa. One partial review that has occurred is *Kitmap – A Stocktake of Research Infrastructure in Aotearoa New Zealand’s Government Research Organisations* (MBIE, 2022b). Notable research infrastructure in that

review included laboratories, field sites, physical and digital collections, computational resources, research vessels and networks of monitoring sensors (MBIE, 2022b). Access was possible to the infrastructure, but this was usually dependent upon policies, time and capacity available, and fees were usually charged for usage of the infrastructure.

Distinct from data infrastructure, *research infrastructure* refers to facilities that provide resources and services for research communities to conduct research and foster innovation (Directorate-General for Research and Innovation (European Commission), 2018). The Kitmap definition of research infrastructure is “a facility or (virtual) platform that provides the scientific community with resources and services to conduct research in their respective fields” (MBIE, 2022b, p. 14). Research infrastructures “can be single-sited or distributed or an e-infrastructure, and can be part of a national or international network of facilities, or of interconnected scientific instrument networks” (MBIE, 2022b, p. 14). Examples of research infrastructure include:

- major scientific equipment or sets of instruments
- collections, archives or scientific data
- computing systems and communication networks, and
- any other research and innovation infrastructure of a unique nature that is open to external users. (Directorate-General for Research and Innovation (European Commission), 2018)

Research infrastructures can also be used beyond research; for example, for education or public services.

Many institutions make use of international infrastructure. The data shows that across the institutions, most infrastructures are primarily funded by the institutions through internal resources (including overheads and commercial revenue), with a lesser amount receiving direct government funding or funding from research grants (workshops). There are occasional instances of co-investment in infrastructures between institutions or with universities or private companies, though these appear to be the exception rather than the rule (MBIE, 2022b).

Enabling the research infrastructure in Aotearoa will require a fairer distribution of design and access across scales and needs. Furthermore, it would improve with a commons model of ownership, with values of respect for prosperity and wellbeing.

STORAGE

The archival and storage spaces are a central part of a data infrastructure. The vast majority of data is stored at an institutional level. Some of the storage is locally stored while other institutions make use of international cloud-based storage such as from Microsoft and Google. Some are using specialty scientific data storage infrastructure.

As data is housed in various places, it is not easy to quantify what data has been collected and the appropriate ways to preserve the data. There are some best practices guidelines for storage and preservation (Archives New Zealand, 2023c), but these have not been socialised throughout Aotearoa.

There are limited large-scale datahouse storage facilities in Aotearoa. At present there is one provider, Catalyst Cloud (Catalyst Cloud, n.d.). As such, there is reliance on international commercial corporations, and especially American multinationals. This dependency on foreign-owned companies and offshoring of data has legal, ethical and economic implications. The offshoring of Māori research data is at odds with

the MDSov Principles and MDGov Model. A recent paper on MDSov and offshoring Māori data argued that MDSov requirements must be central to decision-making, particularly with regard to offshoring and procurement (Kukutai et al., 2022). It advocated a wider suite of storage options including government investment in locally hosted solutions.

There are increasing pressures on data storage infrastructures. The TRUST Principles (Transparency, Responsibility, User focus, Sustainability and Technology; Lin et al., 2020) guide the use of repositories and ensure that they are fit for purpose. Given all these emerging drivers, these Principles ensures the safe collection of data and ensures benefit-sharing to not only other researchers and relevant communities.

DATA SOVEREIGNTY AND JURISDICTION

Data sovereignty (as distinct from Indigenous or Māori data sovereignty) is the idea that data is subject to the laws and governance structures of the nation where the data is collected. A significant amount of data is traversing the globe for many reasons, including for research or in research-associated activities. Sovereignty in regard to the data landscape involves the authoritative claims through domestic institutional arrangements, international regimes and the practices of other states (Hummel et al., 2021). In an ever-globalising world, there is a need to ensure certain safeguards are in place. There is a recognition that sovereign countries need to protect their people, knowledge and national interests (Te Pōkai Tara, 2022, p. 4).

Sovereignty is intrinsically tied into the research data landscape in Aotearoa. Although we collaborate globally, there are key issues of sovereignty, and as part of that, te Tiriti responsibilities. The complexities of sovereignty can occur at all levels, from individual researcher through to national level. Whenever there is a shift of or access to data from Aotearoa to an individual(s) or institution(s) beyond the borders of Aotearoa, there are potential issues of sovereignty. Within an institutional context, sovereignty issues are challenging as institutions own IP and can enable governance (Workshop 2).

Researchers rely on tools and platforms housed outside the jurisdiction of Aotearoa. As a workshop participant noted, “Every tool we use – for example, email, Zoom – has aspects of cloud and offshore considerations, so these considerations transcend all types of tools, platforms, etc., not just of relevance specifically to this discussion.”

A recent legal opinion by a King’s Counsel noted that even if a cloud provider such as Amazon Web Services (AWS) or Microsoft has a base in Aotearoa, the jurisdiction of control is most likely the home jurisdiction of the company; in the examples here, the United States of America (RNZ, 2023). Under such circumstances, sensitive data should be under the control of a cloud provider that is based in Aotearoa. Another potential area of concern is the storing of information and knowledge of taonga species in international databases. There are still limited protections in international databases in terms of partial or full restrictions on data or data sets.

COMPUTING SYSTEMS/POWER

Computing power and systems include levels of hardware, software, processing, informational systems to collect-store and store inputs, computing capabilities, hubs, storage/archive, capabilities. Most researchers in Aotearoa perform research and store their data on personal machines and/or institutional machines. These are often limited in power and performance. In certain fields, the machine’s lack of power and performance inhibits the ability to carry out research activities.

As the detail and requirements of data sets grow, so do the power requirements of computing systems. These requirements are dependent upon the fields and the types of research undertaken. There are few computer systems that have the computer power and capacity needed for bulk computational work. Locally, NeSI (the New Zealand eScience Infrastructure) supports researchers to access and apply such large-scale and powerful systems.

Workshop participants were divided as to whether it is best to rely on ready-made software or platforms or create bespoke software designed for Aotearoa. Bespoke software or platforms can be more expensive, but they can be designed in a way that is specific to the needs of Aotearoa. Conversely, ready-made options are often cheaper, but often have less ability to be adapted or modified to specific needs. The user is also reliant on the owner to maintain the software and the pricing at a level that makes it viable and effective.

DATA REPOSITORIES, ARCHIVES AND COMMONS

There are numerous archives and data repositories in Aotearoa, but there are no regional or national data commons at present. The vast number of repositories are project based. All universities have a repository, with most including provision for research data (either in the same repository or a dedicated research data repository). Stats NZ, through data.govt.nz, currently catalogues 266 data sets from 59 agencies. Most of the data sets are from central government, with a few from Crown Entities, Crown Research Institutes and local government (Booth, 2010, p. 9). MBIE funds the following 26 Nationally Significant Collections and Databases (MBIE, 2023e):

- Marine Invertebrate Collection
- Margot Forde Forage Germplasm Centre
- Adélie Penguin Census Database
- New Zealand Geomagnetic Database
- Water Resources Archive
- Cawthron Institute Culture Collection of Microalgae
- National Earthquake Information Database
- New Zealand Fungarium (Plant Disease Division) – Te Kohinga Hekaheka o Aotearoa
- Te Kohinga Harakeke o Aotearoa – National New Zealand Flax Collection
- National Petrology Reference Collection and the Petlab Database (containing rock, mineral and geoanalytical data)
- Fruit Crops Germplasm
- National Forestry Herbarium database and Xylarium database
- Allan Herbarium and associated databases
- New Zealand Fossil Record File

- National Climate Database
- National Groundwater Monitoring Programme
- Ngā Rauropi Whakaoranga
- New Zealand Volcano Database
- Regional Geological Map Archive and Datafile
- Arable and Vegetable Crop Germplasm
- New Zealand National Paleontological Collection and associated databases
- New Zealand Freshwater Fish Database
- International Collection of Microorganisms from Plants (ICMP) and associated databases
- Land Resource Information System
- New Zealand Arthropod Collection (NZAC) – Ko te Aitanga Pepeke o Aotearoa
- National Vegetation Survey (NVS) databank.

The funding of these databases is under review (MBIE, 2023b). The review is to ensure an efficient and enduring funding model for scientific collections and databases so they can continue to underpin critical science and related activities such as environmental reporting, biosecurity, land-use planning and policy decisions (MBIE, 2023b).

The Aotearoa Genomic Data Repository has been jointly developed by Genomics Aotearoa and NeSI to provide a secure place for the Aotearoa research community to store and share genomic data.¹⁵ Version 1.0 of the repository is up and running, and under ongoing active development and improvement. Another example is Rakeiora, which has been described a “pathfinder” for genomic medicine in Aotearoa. This platform is led in partnership with Māori and responding to iwi-defined health needs. Rakeiora has achieved a world-first level of Indigenous community control or guardianship of data throughout the research process, sustained in perpetuity beyond the initial research. Both of these data platforms are hosted on NeSI’s sovereign Flexible high-performance computing (HPC) community cloud research platform, a platform co-designed to support emerging needs of this type (Genomics Aotearoa, n.d.).

Currently, data sets are spread throughout Aotearoa in varying sizes and formats. Many of these data sets are held on either personal/office systems or sometimes institutional systems. There is no uniformity in how they are stored nor in the policies and procedures surrounding data storage. Another important factor is the limitations of funding of projects. Projects have a limited funding timeline. As such, the data is either lost once the project is finished, or left in an institute storage system or the researcher’s personal system. Due to limitations, researchers often use international databases such as the National Center for Biotechnology Information (NCBI); for example, the National Library of

¹⁵ <https://repo.data.nesi.org.nz/>

Medicine.¹⁶ Many participants in the workshops highlighted the Australian Research Data Commons (ARDC).¹⁷ Such a system overcomes the limitations highlighted by post-funded storage issues, ensuring the data remains available for future research. However, these international sources have their own drawbacks, in particular around sovereignty and MDSov.

Workshop participants noted the ARDC platform that has been established in Australia has improved accessibility to data storage services, as not all institutions have the resources to establish such a platform for themselves. The participants noted the importance of having such a database here, one that could guarantee the long-term preservation of data, but questions were raised such as who would fund it, where it would be stored, and what assurance was there that it would be a protected long-term space.

ARTIFICIAL INTELLIGENCE AND RESEARCH DATA

Artificial intelligence (AI) tools and products are recognisable as a disruptive innovation to many professions, sectors and industries. AI's impact has close interdependencies with and implications for cybersecurity, energy resilience, national sovereignty, social cohesion and democracy. While recent developments in large language models (LLMs) are dominating headlines, this is a high point along a series of recognisable data science developments that are increasing in frequency, and where the road ahead is increasingly seen as uncharted territory.

Artificial intelligence covers “computational methods and techniques that solve problems, make decisions or perform tasks that, if performed by humans, would require thought” (Royal Society, 2019). AI is applied to any decision-making process with an intent to automate human cognition efficiently and consistently through processes and systems that operate at otherwise unattainable scales.

Reviews of data science and AI in the context of research identify data systems and infrastructures as critical ingredients:

Any innovative agenda in data science research and education will depend on a foundation of enabling data infrastructure and useful datasets. Research in data science needs access to sufficiently large and numerous datasets to illuminate and validate results. The datasets must be available for reproducible research and hosted by reliable infrastructure (Berman et al., 2016).

AI is enabled by access to data. To support successful implementation of AI, there is a need for effective digital infrastructure, including data centres and structures for data sharing, that makes AI secure, trusted and accessible, particularly for rural and remote populations. If such essential infrastructure is not carefully and appropriately developed, the advancement of AI and the immense benefits it offers will be diminished. (Walsh et al., 2019)

An increasing array of research methods are used to explore, analyse and codify data in research. A well-established method, machine learning (ML) allows data scientists to provide a computer with a large data set, and teach it to make predictions based on that data. ML, however, depends on human intervention to train models and correct errors.

In recent years the development of ML and AI has exploded with wide use of generative AI large-language models (LLMs) like ChatGPT.

¹⁶ <https://www.ncbi.nlm.nih.gov/>

¹⁷ <https://ardc.edu.au/about-us/>

LLMs depend on deep learning to build their models, based on access to massive data sets and without human intervention. This creates new challenges, including a lack of transparency of the reasoning inherent in the models, and bias and sensitivity contained within the data sets from which the models are built.

The current generation of LLMs are typically trained on data scraped from across global internet websites. This has created an acute sense of urgency for efforts to build open approaches to research data discoverability and sharing, through open science commons and other national infrastructures for research data. Without broad access to the scientific literature, which is often maintained behind paywalls, these models will not represent a balance of scientific knowledge.

Globally, scientific research communities and national research computing and data infrastructure systems are banding together within large scientific language model consortia to address this challenge, recognising that it will only be possible to meet the challenge head on through collective approaches at a global scale.

Generalised global models also have challenges in adequately representing the diversity of communities and their interests. When trained on a global corpus of data, there is limited representation of arcane areas of knowledge or of specific communities, cultures and more diverse systems of knowledge. This reliance on vast data sets can perpetuate bias, especially in contexts where historical inequities exist. Purging biases from algorithms (even if possible) does not necessarily lead to unbiased outcomes, given the inequitable contexts in which such technologies are deployed. This highlights a wider point - that the espoused benefits of AI need to be assessed alongside the risk of deepening inequalities (Royal Society, 2019; Walsh et al., 2019). Effective AI regulation and governance thus need to involve public, communities and stakeholder consultation (Walsh et al., 2019), along with independent, legally empowered oversight (Walsh et al., 2019).

The inherent representational weakness in global models can be partially overcome through models that are trained and tuned to be fit for purpose. Doing so requires AI systems that deeply understand the data, processes and reasoning of specific systems of knowledge. These knowledge systems and their application likely depend on higher degrees of trust and reliability, carrying expectations of high levels of accuracy, trust in the training data and validation of the model's outputs, all with appropriate levels of security, isolation, sovereignty and control. For research data infrastructures and systems in Aotearoa, this means ensuring that there are adequate stocks of data that are representative of our environments, populations and communities, and that such data are appropriately sourced and governed.

The intersection of te reo Māori and ML presents a unique set of ethical challenges (see, for example, Mahelona et al., 2023; McLeod, 2023) and there are significant concerns about data provenance (University of Waikato, 2022), data misuse (Riley, 2023) and the subversion of Māori data sovereignty (Mahelona et al., 2023) and intellectual property rights. In 2022, the AI Institute,¹⁸ TAI AO programme and Tikanga in Technology project teams,^{19, 20} based at the University of Waikato, held a Māori AI wānanga which explored a range of issues including Māori involvement in AI, where AI in Aotearoa should be heading and decolonising algorithms (University of Waikato, 2022). The demand for Māori-led AI approaches was strongly supported. Thus:

Data has been collected by government agencies from a colonial perspective. What we really want to do as Māori is to be in a position where we collect our own data from our own perspectives, because it will look different ... The true essence of Māori data sovereignty is when

¹⁸ <https://ai.waikato.ac.nz/>

¹⁹ <https://taiao.ai/>

²⁰ <https://www.waikato.ac.nz/rangahau/koi-te-mata-punenga-innovation/TinT>

Māori are not accessing data from someone else, as a perspective of someone else's collection, but when we're actually defining, creating, shaping, storing and having authority over our own data, collected from our own unique world view (University of Waikato, 2022).

CONNECTIVE TISSUE

The concept of *connective tissue* encapsulates the mechanisms by which different systems communicate harmoniously and provide for the establishment of universal standards and security frameworks that underpin this connectivity. In essence, connective tissue is the adhesive that binds together the diverse components of our data ecosystem, ensuring data flows efficiently and securely.

One notable example of the connective tissue in Aotearoa's data infrastructure landscape is the *Research and Education Advanced Network New Zealand* (REANNZ). Beyond providing a robust core network infrastructure, REANNZ offers a suite of supplementary services, including Tuakiri, eduGAIN, and Eduroam (MBIE, 2022c, 13).²¹ These services not only facilitate seamless connectivity but also bolster the international reach of New Zealand's research community. Eduroam, in particular, merits special mention, as it grants researchers secure global roaming access to internet services at participating institutions worldwide, using their 'home' log-in procedures and without incurring fees, thus promoting collaboration on a global scale.

Another vital element in our data connective tissue is the *New Zealand eScience Infrastructure* (NeSI), hosted at the University of Auckland. NeSI, a collaborative effort involving numerous organisations,²² serves as the backbone for New Zealand's HPC capabilities and national-level skills development programmes. It boasts two primary platforms: Mahuika, a capacity HPC resource capable of concurrently running numerous small compute jobs, and Māui, a capability HPC resource engineered for executing complex simulations and calculations (MBIE, 2022c, p. 13). Furthermore, NeSI extends its support to institutions by addressing various infrastructure challenges, encompassing RDM and specialised requirements, such as those of Genomics Aotearoa and AgResearch (MBIE, 2022c, p. 15). This collaborative spirit and multifaceted approach exemplify the critical role of connective tissue in our data landscape, ensuring that data flows seamlessly and efficiently across the research and educational sectors in New Zealand.

KitMap serves as an invaluable data repository, offering comprehensive information about the array of infrastructure and instruments available within our data landscape.²³ It acts as an indispensable reference point, aiding researchers and stakeholders in identifying and accessing the requisite tools and resources for their endeavours. In a dynamic and rapidly evolving data environment, Kitmap provides a sense of orientation by cataloguing the diverse infrastructures and instruments at our disposal. This repository empowers researchers, educators and decision-makers to make informed choices, enhancing the utilisation of available resources and driving innovation within the data ecosystem. KitMap is a testament to our commitment to transparency, accessibility and efficiency in managing and navigating the intricate web of data infrastructure in Aotearoa.

²¹ REANNZ provides Tuakiri and eduGAIN as federated identity management systems that allow researchers to access resources (e.g., academic journal subscriptions, log-in to the NeSI infrastructure), using their home institution's identity.

²² NeSI was established in 2010 as a contract-based collaboration between the University of Auckland (as host) and Manaaki Whenua | Landcare Research Limited, the National Institute of Water and Atmospheric Research Limited (NIWA) and the University of Otago (as investing institutions). The University of Canterbury was an investing partner initially but has since left the arrangement.

²³ <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/research-and-data/kitmap/>

FEDERATED SYSTEMS

A *federated database system* (FDBS) is a type of meta-database management system (meta-DBMS) that transparently maps multiple autonomous database systems into a single federated database. The constituent databases are interconnected via a computer network and may be geographically decentralised. Since the constituent database systems remain autonomous, a FDBS is a contrastable alternative to the (sometimes daunting) task of merging several disparate databases. A federated database, or virtual database, is a composite of all the constituent databases in a FDBS. There is no actual data integration in the constituent disparate databases as a result of data federation, but there is a need for accepted standards in order to establish and maintain interoperability.

5. RESEARCH DATA CULTURES

Although the physical and technical elements are what comes to mind when someone thinks of the data landscape, research data cultures were highlighted as a central theme in the many workshops and discussions held as part of this Review. *Research data culture* forms the relational basis to hold the people and teams together and encompasses the behaviours, values, expectations, attitudes and norms of our research communities.

RESEARCH DATA CULTURES

BENEFIT-SHARING

The concept of benefit-sharing has received limited attention within the context of data sharing. While some individuals may be inclined to share data, the extent of such sharing often depends on factors such as the type of data and the circumstances of its collection. Notably, the significance of benefit-sharing has been acknowledged in international agreements such as the WAI 262, the Nagoya Protocol and the CARE Principles.²⁴

A key challenge in fostering data sharing is the inherent competition that exists in the research environment. Data holds immense value and serves as the foundation for complex research projects. However, the competition for limited research funding creates a dilemma where many researchers vie for the same financial resources.

Additionally, it is crucial to address the issue of benefit-sharing with minority communities. Historically, extensive data has been collected from both Māori and Pasifika communities with little reciprocity in terms of benefits returned.

Researchers have expressed concerns and anxieties regarding the notion of *fair and equitable sharing of benefits*. It is imperative to provide clarity on the concept and outline various mechanisms through which benefit-sharing can be realised, thereby alleviating these concerns.

Recognising the inherent value of benefit-sharing, ethical considerations pertaining to the use, sharing and reuse of data are necessary. Pertinent questions in this regard include defining the terms of use or licensing for specific data sets, determining the accessibility of data, and identifying any constraints placed on data access or usage. Ethical guidelines are essential to ensure that the principles of FAIRness and equity are upheld in the sharing and use of data resources.

POLICIES AND INCENTIVES

Policies and incentives are indispensable tools in shaping a vibrant research data culture. They encourage openness, collaboration, accountability and ethical conduct in the management and sharing of research data. By aligning policies and incentives with the principles of responsible data stewardship,

²⁴ The Nagoya Protocol on Access and Benefit-Sharing is an international treaty relating to the governance of biodiversity. The Protocol covers genetic resources and traditional knowledge (TK) associated with genetic resources, as well as the benefits arising from genetic resources where Indigenous and local communities have the established right to grant access to them. (<https://www.cbd.int/abs/infokit/revised/web/factsheet-nagoya-en.pdf>); The CARE principles for Indigenous data governance are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination (<https://www.gida-global.org/care>).

governments and research institutions can contribute significantly to the cultivation of a thriving research data culture within their research communities.

Collaboration between researchers and with the wider economy and society is often hindered by institutional and funding structures. Competition between organisations can get in the way of collaboration, with negative outcomes for stakeholders. Links between New Zealand's firms and research organisations are much weaker than in comparable small-advanced economies. (MBIE, 2022a, p. 9)

Policies and incentives can aid research data culture by promoting data sharing. Institutions and funding bodies often establish guidelines and regulations that mandate or incentivise researchers to share their data. The MBIE Open Research Policy "strongly recommends" sharing of research data (MBIE, 2023a, p. 6) and this is a good first step as such a policy is instrumental in breaking down data silos, enabling the reuse of data for additional research, and fostering collaboration among researchers.

In order for any policy to be realised, appropriate infrastructure, training and support must be in place to make it easy for researchers to comply and make this the default or normative research data culture.

Research funding allocation based on data sharing practices and data management (DM) strategies is a powerful incentive. Government agencies and funding bodies can prioritise projects that demonstrate a commitment to data sharing and responsible DM. Some funders will not fund researchers who have previously failed to comply with their research DM obligations or responsibilities.

In addition, incentives that recognise and reward researchers for adhering to data sharing and responsible DM could become part of a normative research culture. This could include, for example, contribution within researcher metrics and assessment, as well as advancement opportunities.

EXPECTATIONS, NORMS AND CULTURES

There is acknowledgement at agency, institutional and researcher levels of the need to improve the norms and culture in the research field more generally, a sentiment that was also evident in the workshops.

Policies can support this by outlining standards for DM, including data documentation, storage and preservation. These policies can aid compliance by guiding expectations, norms and cultures for researcher best practices in DM, enhancing the quality and accessibility of research data. Setting clear expectations for DM in policy will help create a data culture rooted in accountability and transparency.

Setting expectations in relation to ethical data use and privacy protection is integral to a research data culture that respects individual rights and data ethics. These requirements guide researchers in handling sensitive data, obtaining informed consent and safeguarding privacy. By adhering to ethical data use policies, researchers contribute to a culture of responsible data handling and research integrity. The MBIE Open Research Policy sets out initial expectations for the sector (MBIE, 2023a); however, more direction is required to ensure FAIR and CARE are implemented within Aotearoa.

Any policy development should consider the requirements to support data education and training initiatives. There is little money in the research and innovation sector in Aotearoa, which means collaboration is key. For example, the Council of New Zealand University Libraries (CONZUL) is currently producing an Open Access 101 toolkit aimed at developing researcher skills in relation to OA for all researchers at New Zealand universities, with intentions to share this more widely across the tertiary education sector once the toolkit is completed later this year. Such an approach could be taken to the development of training and educational materials for research data.

ENGAGING COMMUNITIES

COMMUNITY ENGAGEMENT

Community engagement has been notably absent from a substantial portion of research and innovation endeavours. Frequently, the community, or its individual constituents, serve as subjects of research but are seldom integrated into the engagement process itself or the data-generation phases that ensue.

There have been efforts made by certain research teams to establish relationships and involve the Māori community in the engagement process. While there remains substantial progress to be achieved, these endeavours signify a positive step forward.

A pressing necessity exists for communities to actively participate in the entire data lifecycle as well as the governance of data. Mechanisms should be established to facilitate participatory data governance, particularly concerning data that directly affects these communities. For instance, grassroots organisations operating at the community level possess a more comprehensive understanding of community members and their concerns than any single individual. This emerging field warrants dedicated attention within the broader data landscape.

CITIZEN SCIENCE

Citizen science (Science Learning Hub, 2023), although not a novel concept, has seen a gradual uptake within the realm of data science. *Citizen science* involves harnessing the collective efforts of the public and fellow scientists to crowdsource data, thereby contributing to scientific projects. However, its use in data collection and analysis remains limited among researchers. Emerging technologies offer the potential to bolster the accessibility and structure of citizen science methodologies, providing a more comprehensive understanding of communities that often eludes conventional data and research methods.

The adoption of citizen science varies depending on the subject matter and field, with limited formal education and recognition of its significance in research. Nonetheless, citizen science presents a valuable opportunity to tap into the enthusiasm of individuals, including young people and amateur scientists, who harbour a keen interest in specific topics. By facilitating their involvement and support, citizen science brings diverse perspectives to the assessment of human impacts and complements existing data sources.

Citizen science is capable of producing more frequent – albeit potentially noisy – data. Hence the citizen science approach underscores the importance of robust data analysis methods. Addressing biases and distinguishing genuine reports from fake submissions is paramount. Furthermore, the community's access to data is crucial, emphasising the need for co-design and transparent feedback mechanisms.

The application and management of citizen science, however, presents challenges, including around long-term data storage, privacy concerns affecting geolocation data, and issues of data sovereignty. Effective implementation of citizen science, therefore, necessitates training for data collectors, data aggregation to ensure anonymity, and efforts to enhance data set collation and interoperability.

Traditionally, citizen science has played an integral role in Māori mātauranga, involving communities in monitoring environmental and societal changes. In the modern context, citizen science extends beyond individual whānau or hapū, posing potential issues of MDSov. Inclusivity in decision-making processes with Māori representatives is imperative to uphold Māori self-determination and protect Māori data and mātauranga. Incorporating te reo Māori in data collection can further enhance Māori participation. Māori researchers and knowledge holders can provide guidance on engagement strategies and data utilisation.

PEOPLE, SKILLS, EXPERTISE AND WORKFORCE

An often-overlooked cornerstone of the research data landscape is the pivotal role played by the people involved, including their skills and expertise.

SKILLS/EXPERTISE AND SKILL LEARNING

In the research data landscape, several key areas of skill development are emerging as critical components to ethical and successful RDM. The skills, expertise and knowledge required will vary across disciplines, for different roles and at different career stages.

Researchers are increasingly recognising the importance of connecting themselves with their publications, and institutions are actively promoting these connections, often referred to as “reputational toolkits”, to enhance university rankings.

At the core of these connections is personal metadata managed through persistent identifiers (PIDs). PIDs, which are alphanumeric codes, are essential for accurately attributing research to its creators, tracking funding, and strengthening links with data systems. PIDs also play a pivotal role in establishing provenance and attribution, in line with the FAIR and CARE Principles (Brown et al., 2022). By precisely identifying the resources employed in research and the resultant outputs, PIDs significantly contribute to research integrity and reproducibility (Brown et al., 2022).

ORCID is the primary PID system in Australasia, connecting researchers with publications, data sets and awards.

The efficacy of data repositories for future research hinges on the quality of input. Consequently, researchers need easily accessible training resources to learn how to deposit data in a manner that is conducive to further research and adheres to the FAIR and CARE Principles.

Reference has already been made to the requirements to support RDM, including DMPs and DM across the entire research lifecycle. Many researchers outside the traditional “data fields” such as biostatistics and demography assume that the knowledge and skills are not relevant to them. PhD students and ECRs often lack training and support in establishing DMPs, particularly those in fields outside of data-focused subjects.

Initiatives such as “data carpentries” have been established to provide skills and expertise to the community in a relatable format, using familiar examples.²⁵ This approach has the potential to extend its reach both within the broader community and across various levels of the education system.

Given the advancing environment, a comprehensive evaluation is needed to delineate general data skills from those specialised for experts. Certain applications necessitate tailored training, such as the expertise required for managing applications on large-scale HPC infrastructure.

Implementation of training courses should commence at a foundational level and progress to more advanced support. Additionally, a cross-disciplinary approach is essential to ensure that all fields are comfortable with and understand their role within the broader data landscape.

²⁵ See, for example, <https://research-hub.auckland.ac.nz/digital-research-skills/data-carpentry> and <https://www.nesi.org.nz/services/training/overview/software-and-data-carpentry>

A mapping exercise is required to gauge existing skills and expertise, as well as to project future requirements. Establishing a pipeline of skills and expertise is crucial to fortify the stability and security of the data infrastructure. While staff poaching can offer short-term benefits to an institution, it ultimately undermines the stability of both individual institutions and the nation as a whole.

WORKFORCE

The current data research workforce in Aotearoa faces several limitations, including siloed working practices and the rapid adoption of new technologies. While there are pockets of individuals with key skills scattered across various institutions, there is no identifiable, cohesive data workforce in the country. The workforce comprises individuals from diverse disciplines, each varying in their focus on data and expertise levels.

This workforce often operates on an ad hoc basis, with individuals or small teams forming informal networks across institutions. Many have stumbled into this domain through project-related work, and some of their efforts rely on goodwill, extending beyond their formal job descriptions. Additionally, roles in this field tend to evolve over time as skills and knowledge develop and institutional needs change. Ambiguity can arise regarding responsibilities for data-related issues, falling between researchers and professional services.

There is a pressing need for comprehensive staff skills development and capacity building in the data field across Aotearoa. Institutions lack a unified strategy for data systems and processes, necessitating formalised positions to enable structured cross-domain and cross-institutional collaboration. Support roles are crucial for enhancing researchers' effectiveness in the data realm, and stronger links should be forged between government agencies and research organisations to optimise data collection, utilisation and dissemination.

To attract talent and foster professional growth, the data sector must be presented as an appealing employment destination, addressing the current capacity gap and reducing competition within a limited workforce. Incorporating RDM into undergraduate programmes is essential to meet the growing demand for individuals with data knowledge.

Furthermore, there is a scarcity of professionals with expertise in legal education related to data, intellectual property rights and protections. Additional training is imperative to equip individuals with the necessary skills in these domains.

The limited progression of students to PhD candidacy is exacerbated by financial constraints, as many lack support for enrollment in master's and PhD programmes. Scholarships are essential to expanding capacity in the data and data management field.

In addressing the need for capacity building, it is essential to highlight the significant underrepresentation of key demographic groups (McAllister et al., 2019). Specifically, there is a pronounced scarcity of Māori and Pasifika individuals within this field, and few are progressing through educational and developmental pathways. Additionally, individuals with disabilities and those from sex/gender minorities are also underrepresented. These disparities stem not only from equity issues, wherein certain groups lack the same opportunities as others, but also from a systemic problem characterised by limited representation of these minorities from the outset.

To rectify this situation, it is essential to implement programmes aimed at cultivating expertise within these underrepresented groups, drawing upon their unique knowledge and life experiences. Establishing a development pipeline tailored to ECRs from Māori and Pasifika communities is crucial. Such initiatives

will not only promote inclusivity but also enrich the data research landscape by tapping into a diverse range of perspectives and talents.

6. THE WAY FORWARD

SUMMARY

Data and datasets are often described as a core strategic asset for Aotearoa and indispensable for the Government's ambition of being a small nation with an advanced, adaptive and inclusive economy. In the context of our RSI ecosystem, the value of data cannot be overstated. However, as this Review has identified, there are risks and challenges within the wider research data ecosystem that are not being adequately addressed, and potential opportunities and benefits are being missed.

The research science and innovation sector reforms require high levels of collaboration and a culture of co-operation, including the sharing of infrastructure, human capacity and research capabilities.

Collaboration between researchers and with the wider economy and society is often hindered by institutional and funding structures. Competition between organisations can get in the way of collaboration, with negative outcomes for stakeholders. Links between New Zealand's firms and research organisations are much weaker than in comparable small-advanced economies. (MBIE, 2022a, p. 9)

Almost all these challenges either critically depend on, or are substantially enabled by, the collection, production, analysis and use of research data. This Review is timely in order to inform our understanding of where we are now and where we need to go. It is clear that a change of tack is required. The governance and management of research data has been largely ad hoc and inconsistent. Rapid advances in AI and digital technologies, the demand for more accountable and ethical practices, and the growing desire to manage Aotearoa's sovereign interests amid a growing range of domestic and international challenges and opportunities call for a more future-focused approach.

Te Tiriti is a special and important feature of the research data landscape in Aotearoa and its application through data governance and other mechanisms needs to underpin the development of new policies, practices and processes. Tino rangatiratanga is a central component of te Tiriti, with Article 2 guaranteeing Māori authority over their taonga which, in a modern context, includes Māori data. While mātauranga was not within the scope of this review, it is critical that our research data landscape provides appropriate space and resources for mātauranga and its active protection.

This Review has identified a number of broad areas for improvement, including access to data infrastructure, a coherent national approach to data management through common guidelines and procedures, and embedding MDSov and MDGov in the data ecosystem to ensure that Māori sovereignty over their data is protected, and indeed the data sovereignty of Aotearoa more broadly.

These recommendations will critically depend on development of (and change in) research cultures across Aotearoa:

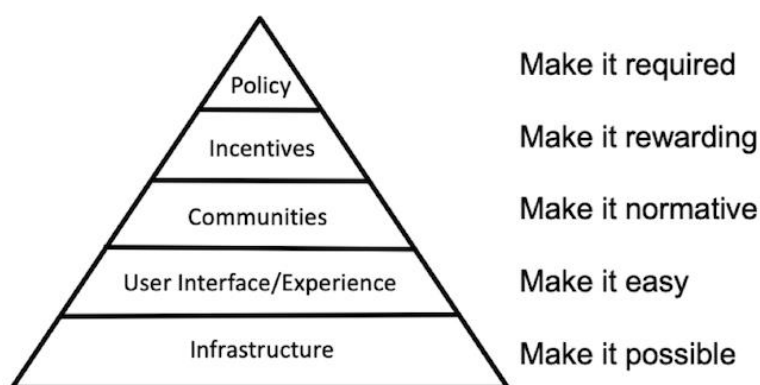
Research culture encompasses the behaviours, values, expectations, attitudes and norms of our research communities. It influences researchers' career paths and determines the way that research is conducted and communicated. (Royal Society, 2023a)

Creating positive research data cultures requires a system of interventions, from a supportive infrastructure to institutional and sector policies that are aligned, integrated and commonly adopted in practice.

The recommendations below do not assume that everything can or should be made open – cultural, ethical and security considerations as well as commercial restrictions will in many cases preclude the application of some open research practices. Open research aims to make things as open as possible – and closed as necessary – and researchers are encouraged to embrace this responsibility and make decisions accordingly across the research lifecycle.

The recommendations are organised under the five headings using Nosek’s Pyramid of Social Change. This structure offers a scaffolded way to systematically approach the work required to develop renewed research data cultures for Aotearoa.²⁶

Figure 3: Nosek’s Pyramid of Social Change



Source: Nosek (2019).

RECOMMENDATIONS

INFRASTRUCTURES – MAKE IT POSSIBLE

1. Clarify which national data repositories are available for researchers in Aotearoa to deposit their data. Consider whether a new de facto repository or a federation of repositories is required for researchers and communities that have no place to store their data securely; for example, because they are not based in a university or research institute. Special attention should be given to locally owned sovereign repositories that can safeguard Māori research data within Aotearoa and give effect to MDSov principles. Some critical questions:
 - How does the infrastructure fit small and diverse research (whatever its scale)?
 - Does the infrastructure support the variety of needs of the various institutions and researchers and the fields, separately and interdisciplinary?
2. Support a review of current research data repositories and databases to assess how research organisations can implement recommendations from the Māori Data Governance Model, and

²⁶ See: <https://www.cos.io/blog/strategy-for-culture-change>

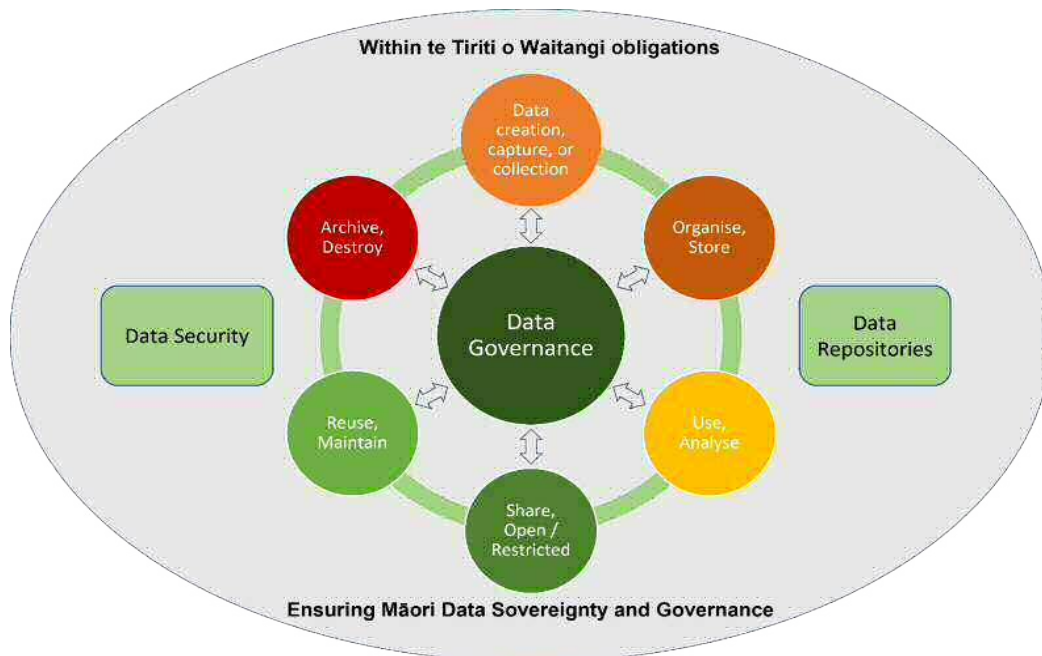
identify further work required. This will require organisations to undertake a stocktake of their Māori research data.

3. Develop a bicultural data ontology. This could build on existing work such as Ngā Upoko Tukutuku (Māori subject headings),²⁷ the work being undertaken by Archives New Zealand (Archives New Zealand, 2018), the model developed by Tauranga City Libraries (Rolleston & Couper, 2021) and the tirohanga Māori mapping of the RSI system by Kanapu (Kanapu, 2023).
4. Improve support for domain- and community-specific infrastructures, including identifying next steps following the MBIE review of nationally significant databases and collections (MBIE, 2023b).
5. Support the work in progress on improving national eResearch infrastructure through MBIE’s Data and Digital Research Institute, including ensuring it takes a holistic, whole-of-system approach to informing and optimising national eResearch infrastructure investment.

USER INTERFACES/EXPERIENCES – MAKE IT EASY

6. Support a community-led effort to examine national and international experiences on RDM across different disciplines/domains. Identify and address issues affecting the (non-)delivery or execution of research data governance, including sustained engagement with stakeholders and dynamic consent. Ultimately the goal is to improve user experiences of data governance cultures, making the interactions represented by the arrows easy and straightforward.

Figure 4: Aotearoa data lifecycle



²⁷ See <https://natlib.govt.nz/librarians/nga-upoko-tukutuku>

7. Provide practical resources to support improving the design of RDM services and engagement with stakeholders across a range of disciplines and help to provide practical guidance to researchers in achieving best practice. Create an RDM101 toolkit that can be used by researchers across the data research sector to gain basic knowledge. This should include support for connecting personal metadata with data and publications. Develop a collaborative toolkit to enhance co-operation among researchers and institutions.

CULTURES: COMMUNITY / BEHAVIOUR / NORMS – MAKE IT NORMATIVE

8. Support the development and implementation of a national strategy for open research, incorporating the FAIR and CARE Principles for Indigenous data.
9. Establish a cross-sector community of practice in RDM tasked with fostering a culture of collaboration among researchers, institutions and the wider community, breaking down silos. This should include those who support RDM but are not researchers.
10. Promote community engagement in the research process, including participatory governance of data. Encourage the use of citizen science as a methodology for data collection. This would include providing support, training and resources; for example, by developing centralised training materials.
11. Develop training and education initiatives to enhance research data culture and promote data sharing. Identify what skills and expertise currently exist and what are required leading into the future. It is expected that skills in data-related legal aspects, such as data rights, IP, and privacy protections, will be flagged. Provide co-ordination of training and upskilling opportunities for researchers and staff in data-related skills; for example, data carpentries.
 - This could be a collaboration of Aotearoa organisations with international initiatives such as the CODATA-RDA Data Schools.²⁸
 - Universities could incorporate RDM into undergraduate and/or postgraduate programmes to ensure that future generations have the knowledge they will need in this area. MDGov and MDSov would be essential components of such training.
 - Create training courses that cover a broad range of data skills and expertise that are tailored to the Aotearoa context, including resources tailored to particular interests and needs.

Through this effort, build a pipeline of expertise by creating programmes to build capacity in underrepresented and poorly served groups. ECRs need more secure employment pathways. Promote the data space as an attractive place for employment and professional growth.

12. Support community efforts to work towards common approaches, processes and guidelines to shape research data culture, including ethical data use and privacy protection.

INCENTIVES – MAKE IT REWARDING

13. Support open research goals (implementing the FAIR and CARE Principles for Indigenous data) through compliance mechanisms within research grants, including for data sharing practices and data management strategies. Include consideration of past open research outcomes within future funding applications (either at individual or institutional level).

²⁸ <https://codata.org/initiatives/data-skills/research-data-science-summer-schools/>

14. Implement incentives that recognise and reward researchers for adhering to data sharing and management practices. This may mean reviewing the importance of current researcher metrics (e.g., through the Performance-Based Research Fund (PBRF)) and shifting the balance further towards evidence of impact. A balance will need to be struck, however, to ensure researchers are not disincentivised from working in areas where data needs to be restricted for cultural or security reasons.
15. Support and encourage ways of celebrating best practice.

POLICIES – MAKE IT REQUIRED

16. Build on MBIE’s Open Research Policy, and move towards mandating RDM as a requirement of researcher funding once other essential elements, such as infrastructure and training, are in place. The policy should maintain the current wording that research data “be made openly available whenever they are not precluded by Indigenous data sovereignty considerations, copyright restrictions, confidentiality requirements, or contractual clauses” (p. 6).
17. Māori governance over Māori research data should be socialised with research organisations and formalised as a requirement of research funding. RDM plans must have MDGov principles and mechanisms embedded throughout.
18. Develop and maintain data policies that adhere to the FAIR and CARE Principles for Indigenous data, ensuring accessibility and usability of research data and MDGov compliance.
19. Set clear expectations for ethical data use and privacy protection through policy implementation.
20. Ensure policies contain guidance that outlines standards for RDM, including documentation, storage and preservation of data.

These recommendations provide a comprehensive framework for shaping government policies and initiatives related to research data in Aotearoa, fostering a culture of responsible data stewardship, collaboration and inclusivity. It is no accident that the culture section has the most recommendations as culture will be critical to successful improvement of the research data landscape in Aotearoa

In implementing these recommendations, it is important to remember that te Tiriti principles sit at the centre of the research data landscape in Aotearoa.

BIBLIOGRAPHY

- Alliance Research Data Management Working Group. (2022). *The current state of research data management in Canada: An update to the LCDRI data management position paper*. https://alliancecan.ca/sites/default/files/2022-03/rdm_current_state_report-1_1.pdf
- Archives New Zealand | Whakapā Mai Ki a Mātou. (2018). *Taonga tuku iho*. <https://www.archives.govt.nz/manage-information/updates-for-information-managers/taonga-tuku-iho>
- Archives New Zealand | Whakapā Mai Ki a Mātou. (2023). *Metadata*. <https://www.archives.govt.nz/manage-information/how-to-manage-your-information/implementation/metadata>
- Armstrong, L. (2023). *Research data classification standard*. ResearchHub. <https://research-hub.auckland.ac.nz/>
- Armstrong, L., Flaherty, B., & Kearns, N. (2021). *Waipapa Taumata Rau Research Data Management Capability Maturity Model v1.2 Research Data Management Project*. Waipapa Taumata Rau | The University of Auckland. <https://doi.org/10.17608/k6.auckland.16993660.v1>
- Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., & Stall, S. (2020). The FAIR Data Maturity Model: An approach to harmonise FAIR assessments. *Data Science Journal*, 19(41), 1–7. <https://doi.org/10.5334/dsj-2020-041>
- Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Hailpern, B., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. (2016). *Realizing the potential of data science: Final report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group*. <chrome-extension://efaidnbmnncnibpcajpcglclefindmkaj/https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf>
- Booth, K. (2010). *New Zealand moves to embrace PSI re-use and open data* (Topic Report No. 15). European Public Sector Information Platform. <https://docs.niwa.co.nz/library/public/TREPSIP15.pdf>
- Bowen, J., & Hinze, A. (2022). Participatory data design: Managing data sovereignty in IoT solutions. *Interacting with Computers*, 34(2), 60–71. <https://doi.org/10.1093/iwc/iwac031>
- Broughton, D., & McBreen, K. (2015). Mātauranga Māori, tino rangatiratanga and the future of New Zealand science. *Journal of the Royal Society of New Zealand*, 45(2), 83–88. <https://doi.org/10.1080/03036758.2015.1011171>
- Brown, J., Jones, P., Meadows, A., & Murphy, F. (2022). *Incentives to invest in identifiers: A cost-benefit analysis of persistent identifiers in Australian research systems*. Zenodo. <https://doi.org/10.5281/ZENODO.7100578>
- Bycroft, C., Miller, S., Gath, M., Matheson-Dunning, N., Simpson, K., & Das, S. (2021). *The quality of administrative data for census variables: Strengths, limitations, and opportunities*. Stats NZ | Tauranga Aotearoa. <https://www.stats.govt.nz/research/the-quality-of-administrative-data-for-census-variables-strengths-limitations-and-opportunities>
- Cabinet Office. (2019). *Te Tiriti o Waitangi | Treaty of Waitangi guidance*. New Zealand Government. <https://www.dPMC.govt.nz/sites/default/files/2019-10/CO%2019%20%285%29%20Treaty%20of%20Waitangi%20Guidance%20for%20Agencies.pdf>

- Cannataci, J. A. (2018). *Right to privacy* (A/73/438). <https://undocs.org/A/73/438>
- Cannataci, J. A. (2020). *Report of the Special Rapporteur on the right to privacy* (A/HRC/43/52). <https://undocs.org/A/HRC/43/52>
- Carroll, S. R., Edit, H., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, 8(108). <https://doi.org/10.1038/s41597-021-00892-0>
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous data governance. *Data Science Journal*, 19(1), 43–43. <https://doi.org/10.5334/dsj-2020-043>
- Carroll, S. R., Rodriguez-Lonebear, D., & Martinez, A. (2019). Indigenous Data Governance: Strategies from United States Native Nations. *Data Science Journal*, 18, 31. <https://doi.org/10.5334/dsj-2019-031>
- Catalyst Cloud. (n.d.). *About us—Aotearoa’s own cloud provider*. Retrieved September 28, 2023, from <https://catalystcloud.nz/about/about-us/>
- Cooke, R. (Sir). (1990). Introduction to commemorative issue to mark the 150th anniversary of the Treaty of Waitangi. *New Zealand Universities Law Review*, 14, 1–8. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2234518
- Council of New Zealand University Libraries (CONZUL). (2016). *Ownership and licencing of research data* [Discussion paper]. Te Pōkai Tara | Universities New Zealand. <https://www.universitiesnz.ac.nz/latest-news-and-publications/ownership-and-licencing-research-data-discussion-paper-july-2016-conzul>
- data.govt.nz. (2023). *New Zealand data and information management principles* Stats NZ | Tauranga Aotearoa. <https://www.data.govt.nz/toolkit/policies/new-zealand-data-and-information-management-principles/>
- Data.NSW. (n.d.). *Data management life cycle*. NSW Government. Retrieved February 2, 2023, from <https://data.nsw.gov.au/IDMF/data-management-and-practice/data-management-life-cycle>
- DARE UK. (2021). *Data research infrastructure landscape: A review of the UK data research infrastructure*. <https://dareuk.org.uk/about/>
- Department of the Prime Minister and Cabinet (DPMC). (2023a). *Information held by government*. <https://www.dPMC.govt.nz/our-business-units/cabinet-office/supporting-work-cabinet/cabinet-manual/8-official-information-1>
- Department of the Prime Minister and Cabinet (DPMC). (2023b). *Critical infrastructure Phase 1 consultation*. <https://consultation.dPMC.govt.nz/national-security-group/critical-infrastructure-phase-1-public-consultation/>
- Dietrich, M., & Jones, N. (2018). *Understanding the eResearch Ecosystem in New Zealand*. Workshop Reflection Report eResearch NZ 2018 Conference. <https://www.nesi.org.nz/sites/default/files/media/eScienceFuturesWorkshop-ReflectionReport.pdf>
- Digital.govt.nz. (2019). *Open data*. Stats NZ | Tauranga Aotearoa. <https://www.digital.govt.nz/standards-and-guidance/data/open-data/>
- Department of Justice Canada. (2023, March 10). *United Nations Declaration on the Rights of Indigenous Peoples Act Action Plan*. Department of Justice - Government of Canada. <https://www.justice.gc.ca/eng/declaration/ap-pa/ah/index.html>

- Directorate-General for Research and Innovation (European Commission). (2018). *European research infrastructures*. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures_en
- Directorate-General for Research and Innovation (European Commission). (2021). *Horizon Europe, open science : Early knowledge and data sharing, and open collaboration*. <https://data.europa.eu/doi/10.2777/18252>
- Dobson, D., & Fernandez, A. (2023). IDSoV and the silent data revolution: Indigenous Peoples and the decentralized building blocks of web3. *Frontiers in Research Metrics and Analytics*, 8, 1160566. <https://www.frontiersin.org/articles/10.3389/frma.2023.1160566>
- European Centre for Medium-Range Weather Forecasts (ECMWF). (n.d.). *Destination Earth*. Retrieved September 28, 2023, from <https://stories.ecmwf.int/destination-earth/>
- Finkel, A. (2017). *2016 National Research Infrastructure Roadmap* (ED16/005828). Department of Education and Training, Australian Government. <https://www.education.gov.au/national-research-infrastructure/resources/2016-national-research-infrastructure-roadmap>
- FNIGC. (2020). *Introducing A First Nations Data Governance Strategy*. The First Nations Information Governance Centre. <https://fnigc.ca/news/introducing-a-first-nations-data-governance-strategy/#:~:text=Available%20now%20on%20FNIGC.ca,First%20Nations%20people%20and%20communities.>
- Garba, I., Sterling, R., Plevel, R. M., Carson, W., Cordova-Marks, F. M., Cummins, J., Curley, C., David-Chavez, D., Fernandez, A., Hiraldo, D., Hiratsuka, V., Hudson, M., Jager, M., Jennings, L. L., Martinez, A., Yracheta, J., Garrison, N. A., & Carroll, S. R. (2023, in press). Indigenous Peoples and research: Self-determination in research governance. *Frontiers in Research Metrics and Analytics*. <https://doi.org/10.3389/frma.2023.1272318>
- Genomics Aotearoa. (n.d.). *Rakeiora: A pathfinder for genomic medicine in Aotearoa/New Zealand*. Retrieved September 28, 2023, from <https://www.genomics-aotearoa.org.nz/our-work/health-projects/rakeiora-pathfinder-genomic-medicine>
- Global Indigenous Data Alliance (GIDA). (2019). *CARE Principles for Indigenous data governance*. <https://www.gida-global.org/care>
- Global Indigenous Data Alliance (GIDA). (2023). *Indigenous Peoples' rights in data*. <https://www.gida-global.org/data-rights>
- Global Open Research Commons Interest Group (GORC IG). (2023). *Typology and definitions*. Research Data Alliance. <https://docs.google.com/document/d/19222GitDg3tcrLsDUqg9Z7GLwPELaS-XHEbN6796Tvc/edit?usp=sharing>
- Greaves, L. M., Lindsay Latimer, C., Muriwai, E., Moore, C., Li, E., Sporle, A., Clark, T. C., & Milne, B. J. (2023). Māori and the Integrated Data Infrastructure: An assessment of the data system and suggestions to realise Māori data aspirations [Te Māori me te Integrated Data Infrastructure: he aromatawai i te pūnaha raraunga me ngā marohitanga e poipoia ai ngā wawata raraunga Māori]. *Journal of the Royal Society of New Zealand*, 1–17. <https://doi.org/10.1080/03036758.2022.2154368>
- Harmsworth, G. R., & Awatere, S. (2013). Indigenous māori knowledge and perspectives of ecosystems. In J. Dymond (Ed.), *Ecosystem services in New Zealand: Conditions and trends*. Manaaki Whenua Press.
- Hudson, M., Anderson, T., Dewes, T. K., Temara, P., Whaanga, H., & Roa, T. (2018). “He Matapihi ki te Mana Raraunga” – Conceptualising Big Data through a Māori lens. In H. Whaanga, T. T. Keegan &

- M. Apperley (Eds.), *He Whare Hangarau Māori – Language, culture & technology* [Ebook], pp. 62–71. https://www.waikato.ac.nz/_data/assets/pdf_file/0006/394908/chapter11.pdf
- Hudson, M., Garrison, N. A., Sterling, R., Caron, N. R., Fox, K., Yracheta, J., Anderson, J., Wilcox, P., Arbour, L., Brown, A., Taulii, M., Kukutai, T., Haring, R., Te Aika, B., Baynam, G. S., Dearden, P. K., Chagné, D., Malhi, R. S., Garba, I., ... Carroll, S. R. (2020). Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics*, 21, 377–384. <https://doi.org/10.1038/s41576-020-0228-x>
- Hummel, P., Braun, M., Tretter, M., & Dabrock, P. (2021). Data sovereignty: A review. *Big Data & Society*, 8(1), 2053951720982012. <https://doi.org/10.1177/2053951720982012>
- Institute of Electrical and Electronics Engineers (IEEE). (2020). *Recommended practice for provenance of Indigenous Peoples' data*. IEEE Standards Association. <https://standards.ieee.org>
- Jackson, M. (2018). In the end “The hope of decolonization”. In E. A. McKinley & L. T. Smith (Eds.), *Handbook of Indigenous education*. Springer. https://doi.org/10.1007/978-981-10-1839-8_59-1
- Jennings, L., Anderson, T., Martinez, A., Sterling, R., Chavex, D. D., Garba, I., Hudson, M., Garrison, N. A., & Carroll, S. R. (2023). Applying the ‘CARE Principles for Indigenous data governance’ to ecology and biodiversity research [online]. *Nature Ecology Evolution*. <https://doi.org/10.1038/s41559-023-02161-2>
- Kamaly, T. (2022). *The importance of data lifecycle management (DLM) and best practices*. IEEE Computer Society. <https://www.computer.org/publications/tech-news/trends/the-importance-of-data-lifecycle-management/>
- Kanapu. (2023). *Home* [Webpage]. <https://kanapu.maori.nz/>
- Kukutai, T., Campbell-Kamariera, K., Mead, A., Mikaere, K., Moses, C., Whitehead, J. & Cormack, D. (2023a). *Māori Data Governance Model*. Te Kāhui Raraunga. https://www.kahuiraraunga.io/files/ugd/b8e45c_803c03ffe532414183afcd8b9ced10dc.pdf
- Kukutai, T., Cassim, S., Clark, V., Jones, N., Mika, J., Morar, R., Muru-Lanning, M., Pouwhare, R., Teague, V., Huria, L. T., Watts, D., & Sterling, R. (2023b). *Māori data sovereignty and privacy* [Tikanga in Technology discussion paper]. Te Ngira Institute for Population Research. https://tengira.waikato.ac.nz/_data/assets/pdf_file/0005/947444/MDSov-and-Privacy_20March2023.pdf
- Kukutai, T., Clark, V., Culnane, C., & Teague, V. (2022). *Māori data sovereignty and offshoring Māori data*. https://www.kahuiraraunga.io/files/ugd/b8e45c_c035c550c8244c70a1025cd90a97298e.pdf
- Kukutai, T., & Cormack, D. (2019). Mana motuhake ā-raraunga: Datafication and social science research in Aotearoa [online]. *Kōtuitui: New Zealand Journal of Social Sciences*, 14(2), 201–208. <https://doi.org/10.1080/1177083X.2019.1648304>
- Kukutai, T., McIntosh, T., Durie, M., Boulton, A., Foster, M., Hutchings, J., Mark-Shadbolt, M., Barnes, H. M., Moko-Mead, T. T., Paine, S.-J., Pitama, S., & Ruru, J. (2021). *Te Pūtahitanga: A Tiriti-led science-policy approach for Aotearoa New Zealand*. <http://www.maramatanga.co.nz/publication/te-p-tahitanga-tiriti-led-science-policy-approach-aotearoa-new-zealand>
- Kukutai, T., & Taylor, J. (Eds.). (2016). *Indigenous Data Sovereignty* (1st ed.). ANU Press. <https://doi.org/10.22459/CAEPR38.11.2016>

- Library Research Services. (2023). *Research data management*. Victoria University of Wellington | Te Herenga Waka. <https://libguides.victoria.ac.nz/research-data-management/research-data-management>
- Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, 7(1), Article 1. <https://doi.org/10.1038/s41597-020-0486-7>
- Local Contexts. (2023). *Grounding Indigenous rights*. <https://localcontexts.org/>
- Mahelona, K., Leoni, G., Duncan, S., & Thompson, M. (2023). Open AI's Whisper is another case study in colonisation. <https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/>
- Manaaki Whenua | Landcare Research. (2023). *Systematics collections data*. <https://www.landcareresearch.co.nz/tools-and-resources/collections/allan-herbarium/systematics-collections-data/>
- McAllister, T. G., Kidman, J., Rowley, O., & Theodore, R. F. (2019). Why isn't my professor Māori? A snapshot of the academic workforce in New Zealand universities. *MAI Journal: A New Zealand Journal of Indigenous Scholarship*, 8(2). <https://doi.org/10.20507/MAIJournal.2019.8.2.10>
- McLeod, L. (2023). How will Chat GPT impact te reo Māori? Data sovereignty experts weigh in. *Radio New Zealand*. <https://www.rnz.co.nz/news/te-manu-korihī/491925/how-will-chatgpt-impact-te-reo-maori-data-sovereignty-experts-weigh-in>
- Mead, A. T. P., Smith, V., Ataria, J., Wilson-Leahy, T. P., Ogilvie, S., Watene-Rawiri, E., Potter, H., & Shadbolt, M. (2022). *When the Crown controls mātauranga: A report on a survey of Crown policies, programmes, legislation, funding, and impact assessment relating to mātauranga Māori*. Biological Heritage National Science Challenge. https://bioheritage.nz/wp-content/uploads/2019/04/When-the-Crown-controls-matauranga_Full.pdf
- Medyckyj-Scott, D., Stock, K., Gibb, R., Gahegan, M., Dzierzon, H., Schmidt, J., & Collins, A. (2016). *Our Land and Water National Science Challenge: A data ecosystem for land and water data to achieve the challenge mission*. https://www.landcareresearch.co.nz/assets/Tools-And-Resources/Data/NLRC-outputs/Our_Land_and_Water_Data_Ecosystem_White_Paper.pdf
<https://www.journal.mai.ac.nz/sites/default/files/Pages%20103%20-%20116.pdf>
- Mika, J., Hudson, M., & Kusabs, N. (2023). Indigenous business data and Indigenous data sovereignty: Challenges and opportunities. *Journal of Intellectual Property & Entertainment Law*, 12, 427–449. <https://jipel.law.nyu.edu/wp-content/uploads/2023/05/JIPEL-Mika-Special-Issue-2023.pdf>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2018). *Open access to weather data review*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/research-and-data/open-access-to-weather-data-review/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2019). *NZ RIS and data management* (NZ RIS 4670). <https://www.mbie.govt.nz/dmsdocument/5733-nzris-and-data-management>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2022a). *Te Ara Paerangi Future Pathways White Paper 2022*. <https://www.mbie.govt.nz/assets/te-ara-paerangi-future-pathways-white-paper-2022.pdf>

- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2022b). *Kitmap—A stocktake of research infrastructure in Aotearoa New Zealand’s Government Research Organisations*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/research-and-data/kitmap/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2022c). *Infrastructure and related services to support New Zealand’s eResearch future*. <https://www.mbie.govt.nz/dmsdocument/25829-infrastructure-and-related-services-to-support-new-zealands-eresearch-future>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2023a). *Kaupapahere Rangahau Tuwhera—Open Research Policy*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/agencies-policies-and-budget-initiatives/open-research-policy/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2023b). *Review of scientific collections and databases*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/funding-information-and-opportunities/investment-funds/strategic-science-investment-fund/funded-infrastructure/review-of-scientific-collections-and-databases/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2023c). *Establishing national research priorities*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/agencies-policies-and-budget-initiatives/te-ara-paerangi-future-pathways/establishing-national-research-priorities/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2023d). *Te Ara Paerangi statement: Embedding te Tiriti o Waitangi*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/agencies-policies-and-budget-initiatives/te-ara-paerangi-future-pathways/embedding-te-tiriti-o-waitangi/te-ara-paerangi-statement-embedding-te-tiriti-o-waitangi/>
- Ministry of Business, Innovation & Employment | Hīkina Whakatutuki (MBIE). (2023e). *Nationally significant collections and databases*. <https://www.mbie.govt.nz/science-and-technology/science-and-innovation/funding-information-and-opportunities/investment-funds/strategic-science-investment-fund/funded-infrastructure/nationally-significant-collections-and-databases/>
- Morris, A. (2023). The dilemma of digital colonialism unmasking facial recognition technology and data sovereignty in Aotearoa New Zealand. *Policy Quarterly*, 19(1), 80–87.
- National Library | Te Puna Mātauranga o Aotearoa. (2023). *New Zealand digital object identifiers*. <https://natlib.govt.nz/collections/digital-preservation/new-zealand-digital-object-identifiers>
- Nosek, B. (2019). *Strategy for culture change*. <https://www.cos.io/blog/strategy-for-culture-change>
- Oliver, G., Lilley, S., Cranefield, J., & Lewellen, M. (2022). Implementing Indigenous data sovereignty: Insights from legislative reform in Aotearoa New Zealand. *Proceedings of the Association for Information Science and Technology*, 59(1), 474–478. <https://doi.org/10.1002/pr2.655>
- Oliveira, M. I. S., & Lóscio, B. F. (2018). What is a data ecosystem? *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1–9. <https://doi.org/10.1145/3209281.3209335>
- Prehn, J., Taitingfong, R., Rowe, R., Garba, I., Price, C., Hudson, M., Kukutai, T. & Carroll, S. R. (2023). *Indigenous data governance and universities communiqué*. Global Indigenous Data Alliance. <https://www.gida-global.org/gidacommunique>

- RAiD. (2022). *Research activity identifier (RAiD)*. <https://raid.org/>
- Rauika Māngai. (2020). *A Guide to Vision Mātauranga: Lessons from Māori Voices in the New Zealand Science Sector*. Wellington, NZ: Rauika Māngai. https://www.maramatanga.co.nz/sites/default/files/Rauika%20Māngai_A%20Guide%20to%20Vision%20Mātauranga_FINAL.pdf
- Rauika Māngai. (2022). *A WAI 262 best practice guide for science partnerships with kaitiaki for research involving taonga. Lessons from Māori voices in the New Zealand science sector*. <https://irp.cdn-website.com/855a29e4/files/uploaded/Wai262-Report-Rauika-Māngai-1.pdf>
- REANNZ, NeSI, & NZGL. (2016). *eResearch 2020: National research data programme: The case for research data*.
- Research Data Alliance. (2021). *GORC International Model WG*. <https://www.rd-alliance.org/groups/gorc-international-model-wg>
- Research Data Alliance International Indigenous Data Sovereignty Interest Group (RDA IG). (2019). *CARE Principles for Indigenous Data Governance*. The Global Indigenous Data Alliance (GIDA). <https://www.gida-global.org/care>
- Research Data Alliance International Indigenous Data Sovereignty Interest Group (2019). *CARE Principles for Indigenous data governance*. The Global Indigenous Data Alliance (GIDA). <https://www.gida-global.org/care>
- Riley, N. (2023). *Māori data is a taonga*. E-Tangata. <https://e-tangata.co.nz/comment-and-analysis/maori-data-is-a-taonga/>
- RNZ. (2023, March 29). *Should official government data be hosted by overseas cloud providers?* <https://www.rnz.co.nz/national/programmes/ninetoon/audio/2018883758/should-official-govt-data-be-hosted-by-overseas-cloud-providers>
- Rolleston, E, & Couper, H. (2021). *Waka Hourua, a model: Recognising and integrating Mātauranga Māori into the practices of Tauranga City Libraries, a case study*. <https://paekoroki.tauranga.govt.nz/nodes/view/53930>
- ROR. (n.d.). *Research organization registry (ROR)*. <https://ror.org/>
- Royal Society. (2019). *The age of Artificial Intelligence in Aotearoa*. Royal Society Te Apārangi. <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.royalsociety.org.nz/assets/Uploads/Artificial-intelligence-Aotearoa.pdf>
- Royal Society | Te Apārangi. (2023a). *Research culture*. <https://royalsociety.org/topics-policy/projects/research-culture/>
- Ruckstuhl, K. (2023). *Data is a taonga: Aotearoa New Zealand, Māori data sovereignty and implications for protection of treasures*. *Journal of Intellectual Property & Entertainment Law*, 12, 391–412. <https://jipel.law.nyu.edu/data-as-taonga-aotearoa-new-zealand-maori-data-sovereignty-and-implications-for-protection-of-treasures/>
- Saunders, T. (2022). *The future is open: Establishing wider open access for research publications in Aotearoa New Zealand*. Office of the Prime Minister’s Chief Science Advisor | Kaitohutohu Mātanga Pūtaiao Matua ki te Piriāia. <https://www.pmcsa.ac.nz/2022/07/15/the-future-is-open-intern-report-on-open-access-publishing-in-aotearoa/>
- Science for Technological Innovation National Science Challenge, Data Iwi Leaders Group, & Victoria University of Wellington. (2018). *Māori Data Futures: Hui report*. <https://www.sftichallenge.govt.nz/assets/Uploads/Download-PDFs/Maori-Data-Futures-Report-2019.pdf>

- Science Learning Hub | Pokapū Akoranga Pūtaiao. (2023). *Citizen science*.
https://www.sciencelearn.org.nz/citizen_science
- Self-Governing Indigenous Governments Data Steering Committee (SGIG DSC) Members. (2023). *The data lifecycle*.
<https://indigenousdatatoolkit.ca/data-management/the-data-lifecycle/>
- Smith, L. T. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples* (1st edition). ZED Books Ltd.
- Sporle, A., Hudson, M., & West, K. (2020). Indigenous data and policy in Aotearoa New Zealand. In M. Walter, T. Kukutai, S. R. Carroll, & D. Rodriguez-Lonebear (Eds.), *Indigenous data sovereignty and policy* (pp. 62–80).
- Stats NZ. (n.d.). *Research using Stats NZ microdata*. Retrieved September 28, 2023, from
<https://cdm20045.contentdm.oclc.org/digital/collection/p20045coll17/search>
- Stats NZ. (2022). *Integrated Data Infrastructure*. <https://secure.livechatinc.com/>
- Stats NZ. (2023). *Research and Development Survey: 2022*. <https://www.stats.govt.nz/information-releases/research-and-development-survey-2022/>
- Switkowski, Z. (2021). *National Research Infrastructure Roadmap*. Available from
<https://ardc.edu.au/article/2021-national-research-infrastructure-roadmap-released/>
- Taiuru, K., Burch, K., & Finlay-Smiths, S. (2023). Realising the promises of agricultural Big Data through a Māori data sovereignty approach. *New Zealand Economic Papers*, 57(2), 172-178.
<https://doi.org/10.1080/00779954.2022.2147861>
- Te Kāhui Raraunga. (2021). *Iwi data needs*.
https://www.kahuiraraunga.io/files/ugd/b8e45c_499e6dc614cd4aa089fe9344c47701ec.pdf
- Te Mana Raraunga (2017). *Te Mana Raraunga*. Available from: <https://www.temanararaunga.maori.nz/>
- Te Mana Raraunga (2018). *Principles of Māori data sovereignty*.
<https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5bda208b4ae237cd89ee16e9/1541021836126/TMR+Ma%CC%84ori+Data+Sovereignty+Principles+Oct+2018.pdfhttps://www.temanararaunga.maori.nz/>
- Te Pōkai Tara | Universities New Zealand. (2016a). *Feasibility study: National Research Data Registry Project report*. <https://www.universitiesnz.ac.nz/latest-news-and-publications/feasibility-study-national-research-data-registry-project-report-july>
- Te Pōkai Tara | Universities New Zealand. (2016b). *Universities New Zealand becomes signatory to ORCID Joint Statement of Principle*. <https://www.universitiesnz.ac.nz/latest-news-and-publications/universities-nz-becomes-signatory-orcid-joint-statement-principle>
- Te Pōkai Tara | Universities New Zealand. (2022). *Trusted research protective security requirements: Guide for senior university leaders in Aotearoa New Zealand*.
<https://drive.google.com/drive/folders/12rdzVNBKugf04BRrbUfVXGJlqOtAQ76>
- Te Pōkai Tara | Universities New Zealand. (2023). *Council of New Zealand University Libraries (CONZUL)*.
<https://www.universitiesnz.ac.nz/about-universities-new-zealand/unz-committees-and-working-groups/council-new-zealand-university>
- Tsosie, R. (2019). Tribal Data Governance and Informational Privacy: Constructing ‘Indigenous Data Sovereignty’. *Montana Law Review*, 80(2), Article 2.
- University of Auckland | Waipapa Taumata Rau. (2023a). *CARE Principles for Indigenous data*.
<https://research-hub.auckland.ac.nz/>

- University of Auckland | Waipapa Taumata Rau. (2023b). *DOIs for research*. <https://research-hub.auckland.ac.nz/>
- University of Auckland | Waipapa Taumata Rau. (2023c). *Research Data Management Policy*. Retrieved September 27, 2023, from <https://www.auckland.ac.nz/en/about-us/about-the-university/policy-hub/research-innovation/research-data-management/research-data-management-policy.html>
- University of Otago Library. (2016). *Data management planning tool*. University of Otago | Mō Te Whare Wānanga o Ōtākou. <https://www.otago.ac.nz/library/dmp/index.html>
- University of Waikato | Te Whare Wananga o Waikato. (2019). *When machine learning, Twitter and te reo Māori merge*. <https://www.waikato.ac.nz/news-opinion/media/2019/when-machine-learning-twitter-and-te-reo-maori-merge>
- University of Waikato | Te Whare Wananga o Waikato. (2022). *Wānanga works to carve a pathway for Māori Artificial Intelligence*. <https://www.waikato.ac.nz/news-opinion/media/2022/wananga-works-to-carve-a-pathway-for-maori-artificial-intelligence>
- University of Waikato | Te Whare Wananga o Waikato. (2023). *University of Waikato open research position statement*. https://www.waikato.ac.nz/_data/assets/pdf_file/0009/999648/open-research-position-statement_2023.pdf?mc_cid=bd37ca5e5d&mc_eid=UNIQID
- Visionary Analytics, Data Archiving and Networked Services, Digital Curation Centre, & European Future Innovation System. (2022). *European Research Data Landscape: Final report*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2777/3648>
- Waitangi Tribunal. (2011a). *Ko Aotearoa tēnei: A report into claims concerning New Zealand law and policy affecting Māori culture and identity*. (WAI 262, Vol. 1). Available from https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_68356416/KoAotearoaTeneiTT2V01W.pdf
- Waitangi Tribunal. (2021). *Report on the Comprehensive and Progressive Agreement for TransPacific Partnership* (WAI 2522). Available from https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_195473606/Report%20on%20the%20CPTPP%20W.pdf
- Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. (2019). *The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing*. Report for the Australian Council of Learned Academies, www.acola.org. <https://acola.org/hs4-artificial-intelligence-australia/>
- Walter, M., Kukutai, T., Carroll, S. R., & Rodriguez-Lonebear, D. (Eds.). (2021). *Indigenous data sovereignty and policy*. Routledge.
- Wilkinson, M., Amos, H., Morton, L., Flaherty, B., Hearne, S., Lynch, H., Lamond, H., Dewson, N., Kmiec, M., Nicolle, J., Skinner, E.-T., & Elliot, G. (2016a). *Research Data Management Framework Report*. CONZUL Working Group, Te Pūkai Tara | Universities New Zealand. <https://www.universitiesnz.ac.nz/files/CONZUL-RDM%20Framework%20Report%202015%20FINAL.pdf>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016b). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). <https://doi.org/10.1038/sdata.2016.18>

APPENDIX ONE: RESEARCH DATA LANDSCAPE PROJECT PROPOSAL

There will be five sections for the report:

1. What does the Aotearoa research data landscape look like? Includes sections making clear the purpose of the report, why it is needed, who is it for, etc. (see structure of MDGov report). Also a brief overview of the current state of play in Aotearoa including Te Ara Paerangi and open data, the study methodology, definitional issues, etc.
2. A 2-page te Tiriti statement as part of setting the scene with reference to Cabinet guidance, the TAP White Paper te Tiriti statement embedded across the data sector, universities, MBIE policy and CRIs (check legislation and also individual CRI policies).
3. The meaty section – graphic showing identification of the key elements of the data research ecosystem and description of each part, hit the sweet spot between too much and too little detail. This might be the section that also gives a brief summary of how other countries have tried to do this type of exercise (or could be as part of background in section 1).
4. Taking a Tiriti lens, what are the challenges and opportunities?
5. Looking forward – ‘next steps’ type of section, 2–3 pages, while we said we’d do guidance, we can’t really deliver that but can definitely provide some signposts here for issues relating to mātauranga, need to defer to the experts.

Drawing from the example of the Canadian research data management landscape review, we propose to undertake a landscape review of data in research. The review will:

- Define the area of research data including:
 - research data across its lifecycle
 - impact and value of research data
 - open and safe data in research
 - Indigenous data sovereignty
 - digital research infrastructure, and
 - people, skills, expertise and workforce.
- Describe the current national investments and support and look at international benchmarks
- Undertake a national landscape stocktake and present what is operating at a national level identifying research organisations, funding agencies, publishers, research-adjacent organisations as well as third-party service providers and international organisations

We expect this will lead to a follow-on project that will identify challenges and opportunities for the current research data ecosystem. It is anticipated that this will support the next steps of Te Ara Paerangi.

APPENDIX TWO: DEFINING RESEARCH DATA

Defining the term *data* is not as simple as it sounds. Though it is widely used, it is not always defined. Some beginning points for the definition of data can include some of the following from various national and international sources. For example, the *Our Land and Water National Science Challenge: A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission* defines data as:

... opinions, reports, observations, facts, and statistics collected or created for a specific purpose of studying or analysing, gaining understanding and communicating ... [and] modelling and data-related publications. (Medyckyj-Scott et al., 2016, p. 2)

While Wilkinson, Amos, et al. (2016) define data as:

... facts, observations or experiences on which an argument or theory is constructed or tested. Data may be numerical, descriptive, aural or visual. Data may be raw, abstracted or analysed, experimental or observational. Data include but are not limited to: laboratory notebooks, field notebooks, primary research data (including research data in hardcopy or in computer readable form), questionnaires, audio and video recordings, models, photographs, films or test responses. Research collections may include slides, artefacts, specimens and samples. Increasingly these objects are being captured in digital forms via sensor arrays, electronic notes and digital image capture. (p. 56)

Data, both physical and digital, encompasses a huge variety of types of objects such as word processing documents, spreadsheets, database files, charts, graphs, electronic mail, logs, photographs, programming notes and so forth (Medyckyj-Scott et al., 2016, p. 2). There are also modern forms of data; for example, archived websites and social media aren't mentioned in the above definition (Workshop 1). And universities and other institutions have other additional data classification across the three 'types' of data – education, administrative/institutional and research (Workshop 1).

For this Review it is important to ensure that a definition includes or encompasses what it means in a research context. The Canadian report on research data management defined research data as:

Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. ... Research data may be experimental data, observational data, operational data, third-party data, public sector data, monitoring data, processed data, or repurposed data. (Alliance Research Data Management Working Group, 2022, pp. 8–9).

The University of Auckland uses the following definition of research data:

... the evidence that underpins the answer to a research question and can be used to validate findings regardless of its form (e.g., print, digital, or physical). Research data does not include institutional data or primary materials. (Research Data Support Services, 2023)

Specific fields may have their own understandings or definitions of research data and do not need to be expounded here. For example, non-traditional data fields such as creative works also need to be included in the understanding of data (Workshop 1). Furthermore, the understanding of data includes both the raw and processed, qualitative and quantitative, and physical and digital forms of data.

In addition to research data, there is also a need to consider *administrative data*. Administrative data refers to data collected by organisations for the purpose of running and monitoring programmes or services. Administrative data is defined by Stats NZ as:

... data collected by government agencies or private organisations in the course of conducting their business or services. It is data not collected primarily for statistical purposes. Rather, it is collected for operations such as delivering a 'service' (for example, health or education), or legal requirements to register events (for example, births, deaths, and marriages) or as a record of transactions or events (for example, tax payments and overseas travel journeys). The population and data content is defined by the collection organisation and they have primary control of the methods by which the administrative data are captured and processed. As a result, administrative data differs in nature, scope, and quality to data collected directly through the census or surveys, where control of who is asked for what information is in the hands of the statistical agency. To date, most of the admin sources investigated by census transformation are generated by government agencies. (Bycroft et al., 2021)

With regard to research, the *Our Land and Water National Science Challenge: A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission* notes administrative data as:

... such as those arising from project management and contracts, and data in the form of documents such as reports and research publications, are however, considered out of scope. (Medyckyj-Scott et al., 2016, p. 2)

All data – research or administrative – can be classified as public, internal, sensitive and restricted (Armstrong, 2023).

In summary, research data is not easily definable. However, it encompasses a wide range including raw, processed, qualitative and quantitative, physical and digital forms required from which an argument or theory is constructed or tested.

Data, as described above, originates in and is expressed and analysed in various contexts. These contexts include both individuals and collectives, human and non-human, environmental, and other sources. They are gathered by governmental agencies and research institutions. The data and associated research activities vary between these contexts and add complexity in attempting to evaluate the research data landscape in Aotearoa.