The Future of Science is Digital

Te Ara Paerangi seeks to understand what changes are required to the Aotearoa/New Zealand's Research, Science, and Innovation ("RSI") ecosystem to best position it for the future. It explores six potential areas of action:

- Research Priorities
- Institutions
- Funding
- Te Tiriti, Mātauranga Māori, and Māori Aspirations
- The Research Workforce
- National Research Infrastructure

This submission will focus on the last area, National Research Infrastructure. However, it will succinctly comment on other areas when related.

Framing is not ambitious

Te Ara Paerangi describes the National Research Infrastructure as *"laboratories, equipment, and collections and databases [which] are essential inputs into research activities and science services."*--in other words, enabling technologies. The report further lumps these with IT systems and data infrastructure when it suggests that *"there is an opportunity for increased efficiency through co-location or shared use of infrastructure resources, such as IT systems, to make more efficient use of capital investments"*, or when it describes that *"Research infrastructure can also include key data infrastructure. This includes scientific databases, such as those containing weather or environmental data, and invaluable social research data, such as the results of existing cohort studies and the ability to commission new cohort studies. Lack of specific ongoing funding can mean some of this data infrastructure struggles with maintenance over time."*

Later, the report describes how co-location could reduce the cost of the infrastructure. This framework re-enforces the status quo, in which the chronic underfunding of IT systems and data infrastructure means that Aotearoa/New Zealand lingers behind global digital research, rather than leading or developing it.

We consider this approach unambitious because it ignores the eResearch component at its centre. We define eResearch as a new science discipline that combines data science, data engineering, DevOps, machine learning engineering, and related digital science capability. eResearch allows exploring large and traditionally unrelated datasets to expand knowledge beyond what current methods allow. We posit that these practices will continue to permeate through every layer of the scientific research activity, integrating targeted or domain-specific research outputs into a multidisciplinary view of a specific sector. Eventually, this will become a core competency of operationalised science generation, much like it has happened with digital technologies being adopted in every other aspect of the modern workplace (e.g.: we no longer produce reports with typewriters).

Characterising eResearch merely as *"infrastructure"* clearly undervalues its potential thus, resulting in chronic underinvestment and under resourcing. Lack of investment in eResearch capability means pointless duplication of data and methods, unnecessarily redundant capability, increased competition between CRIs and other organisations for the same scare resources (human or otherwise), uncoordinated research, loss of potential innovative knowledge, limited value of isolated datasets, loss of ability to realise economic benefit, inflated costs, inability to attract top international talent, and/or loss of domestic talent. All of these will vastly negate any efficiencies that may be gained by the co-location of the physical IT infrastructure that underpins the practice.

We do recognise, however, that managing eResearch infrastructure (such as High-Performance Computing and research data platforms) requires specialised IT and service delivery skills that are globally in short supply. We also know that digital skills required for researchers to make the most of this specialised infrastructure are not trivial to learn, especially for seasoned researchers who have managed so far without needing them. In this context, some centralisation does appear desirable to maximise the use efficiency of those resources. However, that same centralisation can lead to added bureaucracy that stifles agility and innovation, or inappropriate governance structures that hinder participation. Aotearoa/New Zealand must find a good balance between grouping scarce resources and capability providers, and decentralised decision-making and rapid innovation.

In this submission, we propose the creation of a “National Centre of eResearch Excellence” to concentrate the specialised capability required to drive the digital transformation of science in Aotearoa/New Zealand. We also explore why such a Centre could help create a well-connected and collaborative RSI ecosystem, working for the benefit of Aotearoa/New Zealand and even humanity.

eResearch is the only way forward for science

Data is the cornerstone of research production[1]

Data is already conceptualised as an economic concept[2] that is core to the world beyond the current fourth industrial revolution. Already recognised as such by most actors of the market economy, policymakers are now catching up to acknowledge and build on its pivotal role in other domains, like societal values, sovereignty, ... and research.

Data is not just a key asset for business success; it is also considered the "real currency of good research and scientific knowledge"[3]. In this context, it comes as no surprise that many of the efforts being undertaken by CRIs and other companies have lately been focused on establishing data-driven capabilities leading to in-depth digital transformation that will alter how science is applied and how impact gets realised irreversibly. This capability is the embodiment of eResearch and represents a new scientific discipline that is not limited to data form or subject matter.

Dealing with data in any organisation comes with challenges. Dealing with scientific research data makes these challenges even more difficult:

1. There exists a tension between the CRIs mandate of “carry out research [...] for the benefit of New Zealand” and the requirement to be "financially viable" and operate on commercial lines. Whilst an increasing number of scientific journals now demand that raw datasets are included for

[1] Coined from the recently popular “Data is a factor of production” trend, cited by Lillian Li: Abridged: Data as a factor of production - by Lillian Li (substack.com)
[2] Policy Options for the Data Economy - a Literature Review (cpb.nl)
[3] Data is the Real Currency of Science

publication, often those same datasets are subject to commercially sensitive or industrial or trade secret rules, which prevent the benefits of Open Data and Open Science from being realised.

2. The submission of data during the publication of scientific articles is largely done in an unregulated manner with little or consideration around who and how that data will later be used, or where will it be stored. This issue arises from an absence of a nation-wide strategy and corresponding governance policies around Aotearoa/New Zealand's research data, that would facilitate better publication practices.
3. Dealing with Taonga data can bring challenges unique to Aotearoa/New Zealand from data governance and IT management perspectives, since the western "ownership" and "intellectual property" constructs (around which technology is usually built) do sometimes translate into different or, in some cases, opposite concepts in Te Ao Māori.

Digitalisation of research data can help solve some of these problems. Technology solutions allow for the creation and orchestration of arbitrary rules around data ownership, access controls, permissions, storage, security and kaitiakitanga. But often, those same technological solutions create additional complexity, reducing accessibility or becoming a barrier when the end user of that data (or some other stakeholder) lacks digital literacy. Collaboration can also be hindered when technology stays hidden in each organisation's walled garden because of trade secret, data protection, intellectual property protection, or mere competitive advantage.

A National Centre of eResearch Excellence could overcome some of these challenges by providing centralised capability to CRIs and other research organisations around:

- Co-design of data governance, security, and protection rules and technology solutions
- Socialisation of good practices for data collection, ingestion, storage, retention, archival, and disposal
- Management and maintenance of domain-specific digital tools used for Nationally Significant Databases and Collections

From data to insights

The process for science delivery starts with a hypothesis. Data is then collected and analysed hoping to prove that hypothesis. This "analysis engine" is the core of the RSI machinery, and while several aspects of the current system can be criticised and improved, we believe that in the long term, these will be the main challenges that will need addressing from a national policy perspective:

1. Certain industries in Aotearoa/New Zealand have succeeded thanks to publicly funded research that helped kick-start them (eg: twenty years of government-funded research into breeding, orchard management and post-harvest storage and transport was a major contributor in the commercial success of the kiwifruit sector). Those industries now got to the point where they can continue funding themselves. However, not all sectors are in the same position, partly because of lack of public funding.
2. Policies to access and manage science data differ between institutions. Efforts to standardise data governance policies are not coordinated between CRIs, creating silos that complicate inter-CRI collaboration. The expansion of capabilities is driven internally by each CRI without any consideration or concertation with other academic organisations. As a result, infrastructure

duplication has occurred with not much potential to align and consolidate this infrastructure or the associated policies in the future.

3. Scientists and organisation leaders have identified a need to move from research that describes the behaviour of a system, through to research that predicts how a system will behave, to prescriptive research that attempts to engineer the system to obtain the desired set of behaviours. This is not a new view and indeed is becoming a common framework for this digital journey.[4] It does, however, pose questions around how to best organise the RSI ecosystem for this transition of paradigms to be most effective.

A National Centre of eResearch Excellence could provide centralised capability to CRIs and other research organisations around:

- Providing specialised technology, analysis platforms, and low-level systems and software engineering expertise (eg: High-Performance and Scientific Computing, data analytics platforms, MLops, etc); to maximise the efficiency of IT resource consumption for digital scientific analysis.
- Maintaining a collaboration-based community of practice to foster connectivity, worker mobility, and knowledge transfer across the RSI ecosystem.
- Reporting to Government on the impact that digital science creates for diverse sectors, which will help inform future investment decisions.

Knowledge transfer across RSI [p]layers must be fluid

Pathways to the future describes an NZ RSI ecosystem made up of three layers:

- Disruptive technology (Start-ups)
- Applied science (CRIs / Industry)
- Fundamental science (Universities / CRIs)

As systems become more complex, automated, and large, it is becoming increasingly difficult to translate outcomes of 'old-fashioned' research from universities and CRIs into start-ups and industry. In a small market like NZ, costs of doing so can be prohibitive. Instead, if new research were conducted using digital frameworks already in use by industry and start-ups, the connection between each of those layers would be more fluid —allowing widespread adoption of innovations that were designed to benefit NZ. We believe that CRIs cannot afford to do fundamental science whilst continuing to ignore how industry or start-ups leverage that science to create novel products, or to increase the efficiency of the sectors in which they operate.

The way our fundamental science is done has to change, and we should start operating with a start-up/industrial mentality (e.g.: by leveraging technology to disrupt science, instead of merely using it as an avenue to reduce costs and time to market of new offerings). We have made strides towards this by adopting technology like version control and digital playbooks, but we are yet to adopt a "data engineering" approach that streamlines connections across discrete research areas.

In lay terms, the way this problem manifests is as follows:

---

[4] What Is the Difference Between Descriptive, Predictive and Prescriptive Analytics (bernardmarr.com)

a. Public grants or industry sector demands kick-start research in a specific area to optimise its productivity or find answers to unknowns that could create impact if solved
b. A research project around the problem is planned, proposed, and eventually funded
c. Digital solutions are used through the lifecycle of the project for data capture, data analysis, publication and/or reporting on results, and project management. During this process, novel algorithms, models, or digital services might be conceived that either function as proof-of-concept or small-scale solution to the problem being researched on
d. When the project closes researchers move on. Ongoing maintenance of the digital solution is usually non-existent. In the best of cases, industry might adopt it and further develop it into a product. In most cases this never happens, and the same wheel is later reinvented multiple times.

The difficulty of transitioning between fundamental research and development, to a commercial market-ready product, is colloquially known in business as "the valley of death" problem. We are not suggesting that additional investment in eResearch will solve it, but we do believe that government policy can positively contribute by creating an RSI ecosystem where the friction of knowledge transfer is reduced. Closer alignment between the digital practices of "fundamental science" generators to the ones used in industry and start-ups contexts, would also contribute positively to workforce mobility, with increased competition in the sector driving up working conditions for everyone.

A National Centre of eResearch Excellence could address some of the above, not only by maintaining a robust and collaborative community of practice in the digital science space to foster alignment between digital science and industrial-grade IT engineering; but also, by hosting nationally significant digital databases and collections, which require specialised domain-knowledge to be maintained and administered. For instance, if all kiwifruit germplasm phenotypical and genotypical data were to be hosted centrally, with robust role-based access controls that protect the commercial interest of stakeholders, more organisations could access this data to further research in areas previously inaccessible to them.

Pipeline of talent should flow both ways

Traditionally one gets formal training at university and "real-life" experience at work. Only a small portion of the workforce gets formally retrained to use new digital tools and this is hindering making robust science. We can neither wait for graduates to replace the current workforce, nor is it viable to completely suspend work in order to retrain the workforce in the short term. The beginning of an answer is to fill this gap with a different model that allows for the continued and ongoing upskilling and reskilling of researchers.

Universities and other education providers could assume a role of "perpetual educators" beyond undergraduate and postgrad studies. This already exists in the educational and health sectors. Ideally, every researcher should be able to allocate a set amount of time on a regular basis to upskill themselves in their field, not just to catch up with digital technology but also other areas such as Intellectual Property, Legal, Te Ao Māori, etc.

A National Centre of eResearch Excellence could provide some training on their own, as well as suggest other focus areas that third-party training providers could cover. Training achievements could be tracked by NZQA through micro-credentials[5].

National Research Infrastructure should not compete with market

Centralising infrastructure can help maximise the efficiency of investments by reducing costs, avoiding duplication, and grouping supporting capability; this also applies to digital infrastructure. However, the way this central infrastructure is brought to institutions and set up to respond to research needs in a timely manner will make or break the model.

What works best is a national infrastructure organised around a provider-subscriber business model, complementing or competing with open market offerings that would follow the same procurement process rules. The value-add of a service provider who not only understands the specialist needs of its consumers but can also act as a central hub for collaboration across institutions, simply because it sits outside of them, can be reason enough to be preferred over a neutral commercial competitor. A good example of this is REANNZ, who can compete with other Internet Service Providers in the Aotearoa/New Zealand market by providing several additional services atop of barebones network connectivity (eg: Tuakiri, EduGAIN, eduroam, etc). These services reduce barriers to collaboration, create a standardised offering that researchers become familiar with, and reduce friction during worker mobility.

There is however a risk that centralising specialised capability creates a silo and stifles innovation. A service provider with a captive market can survive being disrupted by novel technologies, which may lead to an ivory tower situation: they do not need to improve nor modernise their offerings because their consumers benefit so much from those highly specific, value-add services, that moving to another provider becomes undesirable. Market competition is the only solution here. Creating incentives for use of shared infrastructure while in parallel allowing RSI organisations to continue consuming open market services when they are more effective, will create tension between the two. This tension can be resolved by the shared infrastructure provider adopting emerging technologies, and maturing and adapting them to the unique needs of the RSI ecosystem. Again, REANNZ here becomes exemplar of this: market technologies such as Virtual Private Networks (VPNs) and Public Cloud Gateways which extend an organisation's private network, are now REANNZ value-add services (eduVPN and Cloud Connect, respectively) which members can readily adopt.

A bimodal eResearch strategy for Aotearoa/New Zealand science

In the domain of Information Technology, *"Bimodal is the practice of managing two separate but coherent styles of work: one focused on predictability; the other on exploration. Mode 1 is optimized for areas that are more predictable and well-understood. It focuses on exploiting what is known, while renovating the legacy environment into a state that is fit for a digital world. Mode 2 is exploratory, experimenting to solve new problems and optimized for areas of uncertainty. These initiatives often begin with a hypothesis that is tested and adapted during a process involving short iterations, potentially adopting a minimum viable product (MVP) approach. Both modes are essential to create substantial value and drive significant organizational change, and neither is static. Marrying a more predictable evolution of products and*

[5] https://www.nzqa.govt.nz/providers-partners/approval-accreditation-and-registration/micro-credentials/

technologies (Mode 1) with the new and innovative (Mode 2) is the essence of an enterprise bimodal capability. Both play an essential role in digital transformation."[6].

eResearch IT infrastructure usually comes in the form of High-Performance Computing facilities and Data Analytics platforms that operate following both traditional system administration and modern software engineering practices such as DevOps. We believe that the provisioning and maintenance of this infrastructure has now matured enough to the point where it can be treated as "Mode 1" described above. It does, however, continue to be a niche and specialised practice, dependent on highly specialised skills which make the offerings hard to commoditise (e.g.: by a public cloud provider). Attempts at commercialising some of this capability (e.g.: "High Performance Computing as a Service") exist in the global market, however the number of players in this space is small, and their products are somewhat generic, as commoditisation tends to cater for the most common denominator. Additionally, one still needs specialist skills to consume these services in a cost and research-effective manner. Just because a researcher can create a cluster in the public cloud, it does not mean that it will be used sensibly!

It is in this market context that CRIs and Universities continue to see success by adopting shared computational facilities like the ones managed by New Zealand eScience Infrastructure (NeSI), a collaboration between University of Auckland, NIWA, the University of Otago, and Manaaki Whenua - Landcare Research. NeSI, however, does not have a Central Government mandate to be the provider of choice for eResearch infrastructure. As a supra-entity, it receives funding and determines strategic priorities based on the coordinated efforts of its constituent partners, with some elements of co-design driven by the needs of their subscribing users. The process for a new CRI or some other organisation to join this partnership not only requires a high-level of buy-in from their own internal governing body, it also needs the other partners to accept it, assuming that their own priorities continue to align. Joining NeSI as a partner can also only be done when their contract is being renegotiated. These factors present a high barrier of entry that hinders participation. In response, NeSI have explored different business models that work around them with various degrees of success, however adoption is not yet uniform across the RSI ecosystem.

While NeSI could comfortably expand their subscriber base and act as Aotearoa/New Zealand's specialised service provider for eResearch infrastructure, not all decisions around use of technology in a science context can nor should be fulfilled by them. Instead, researchers should retain the ability to explore and adopt novel and innovative services from the market for their own needs. This effectively translates to CRIs and other organisations focusing on "Mode 2" styles of work. It is worthwhile reinforcing here that neither of these strategies are static. While organisations like NeSI might rest in their ivory tower, capability would mechanically get transferred between NeSI and CRIs if the former were mandated to maintain a community of practice for eResearch. Over time, "Mode 2" technologies become "Mode 1" and get adopted at National level when there are efficiencies to be realised. Similarly, as Aotearoa/New Zealand eResearch practice evolves, concentrating specialist skills at a National-Provider level could lead to innovative technologies being developed to better satisfy the needs of the local community. These technologies can also later transfer back to market to be further commoditised. We see this constant pendulum between innovation and mass-adoption all the time in the open-source software space.

[6] https://www.gartner.com/en/information-technology/glossary/bimodal

In short, we would like to propose the following high-level action points that would contribute to foster robust eResearch capabilities in Aotearoa/New Zealand's scientific community:

- Detach NeSI from its founding partners and make it into a not-for-profit Crown-owned company or cooperative (on the same vein as REANNZ) to allow it to expand and deliver on its refreshed mandate to all other RSI organisations equally.
- Position NeSI as a National Centre of eResearch Excellence (NCeE), with a mandate to maintain capability and IT infrastructure that covers the following (this list is non-exhaustive):
  - Traditional High-Performance Computing
  - HPC as a Service / "Flexi-HPC" / on-demand HPC
  - Development of eResearch technology that specifically supports the needs of Aotearoa/New Zealand science
  - Provide and maintain a National Data Archive for scientific data that follows FAIR[7] and CARE[8] principles
  - Support for biocultural labelling[9] of material stored within NCeE
  - Hosting of Nationally Significant Databases and Collections (e.g.: Genomics Aotearoa) whenever another organisation would not be more appropriate
  - Provide for data security and data protection best practices (in coordination with CERT NZ)
  - Foster eResearch learning and development, in coordination with training providers such as The Carpentries, Universities, and NZQA
  - Maintain a collaborative community of practice (in coordination with REANNZ for connectivity and authentication services)
- Create disincentives for CRIs to duplicate NCeE's capabilities, and/or incentives for using it. For instance, prioritise funding of research plans that have a focus in collaboration using shared infrastructure.
- Empower CRIs to leverage public cloud services where agility is required
- Seek All of Government agreements with major cloud providers that can help streamline procurement processes
- Negotiate with cloud providers on behalf of NZ RSI for access to training / discounted pricing / etc.

In this model, the National Centre for eResearch Excellence takes the centre stage as the main provider of eResearch infrastructure and digital science practice capabilities, empowering and sustaining a community that will speed up and enhance collaboration across the entire RSI ecosystem. We can think of it as the metropolis around which other RSI organisations position themselves, acting as the "cities" where day-to-day research happens; while REANNZ continues providing the communication and

---

[7] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[8] The CARE Principles, Research Data Alliance, International Indigenous Data Sovereignty Interest Group. (September 2019) "CARE Principles for Indigenous Data Governance" - The Global Indigenous Data Alliance, https://www.gida-global.org

[9] https://www.enrich-hub.org/bc-labels

information highway between them, as well as providing connectivity to public cloud providers when rapidly changing business needs require using them.

Investing in eResearch as a core competency and making it a defining characteristic of our RSI ecosystem, will lead to a well-connected and digitally savvy research workforce. Aotearoa/New Zealand science will thrive as a result.

Authors:

Cecilia Deng, Team Leader Bioinformatics Engineering
Eric Burgueño, High-Performance Computing Services Manager
Guilhem Morra, Science Systems Analyst
Melissa Broussard, Scientist
Peter McAtee, Scientist
Simon Deroles, Computational Research Operations Leader

The views expressed in this submission are our own, and may not necessarily reflect the views of our employer