Submission on the National Research Infrastructure Roadmap Capability Issues Paper

from the National Computational Infrastructure

9 September 2016

Name	Professor Lindsay Botten
Title/role	Director
Organisation	**National Computational Infrastructure (NCI)**

National Computational Infrastructure
Roadmap Issues Paper Response

1. SUMMARY (in response to opportunity to provide 'Other Comments')

The National Computational Infrastructure (NCI) welcomes the opportunity to provide input on the *Capability Issues Paper*, and to contribute to the development of the 2016 National Research Infrastructure Roadmap. Our submission focuses largely on the national e-infrastructure, namely the high-performance computing (Sec. 10), and the Data for Research and Discoverability (Sec. 11).

Over many years, investments in high-performance computing (HPC) globally have delivered astounding benefits in all advanced economies—underpinning economic competitiveness and providing a foundation for the solution of grand challenge problems, and the substantial advancement of research in every field of modern science and industry application.

Today's HPC environment has evolved to encompass the needs of big data (processing, analysis, data mining, machine learning), in addition to its traditional role of computational modelling and simulation. The contemporary environment comprises a tightly-integrated, high-performance infrastructure able to handle the computational and data-intensive workflows of today's research, together with expertise in computational science, data science and data management. This *fusion of big computation, big data*[1] *and expertise* is today referred to as *advanced computing*. Pursuing such a direction is being advocated strongly in the USA, Europe, the UK and many other advanced economies, as it is the converged support delivered by advanced computing which is essential for the research domains that stand to deliver the greatest contribution to an innovation economy.

In Australia, NCI is amongst the forefront of these developments. Advanced computing is already embedded in the Australian economy and its research and innovation system, playing a vital role in empowering social, economic, and environmental outcomes. Exemplifying this, through NCI, are:

- *Adaptation to Extreme Weather, and Regional Climate Impacts* – next-gen weather/climate models developed by BoM, CSIRO, and an ARC Centre of Excellence on HPC research systems provide essential tools for policymakers; deliver benefits to industry by improving safety and reducing losses (up to billions of dollars) from extreme weather events through better prediction; and underpin better seasonal prediction, leading to potentially billion dollar benefits to the agricultural sector.
- *Environmental and Agricultural Policy* – the availability of Australia's reference environmental and geoscience datasets via NCI's environmental data interoperability platform provides a unique capability for innovation—such as the Australian Geoscience Data Cube. This highly-processed, petabyte archive of 35 years of Landsat satellite earth observation data enables analysis and prediction of flood risk, land degradation, and aquifer recharging, to inform environmental and economic policy, and underpin the environmental data products made available to industry.
- *Advanced Extraction Techniques for Resources*—HPC modelling has translated university IP for determining rock porosity, into a successful Australian start-up servicing the oil and gas sector, and which sold to a US multinational for US$76M in a trade sale in 2014.
- *Health and Medical Research*—vital high-performance analysis[2] of human genomes is advancing our understanding of cancers, autoimmune and other diseases—leading to personalised medicine in the future, benefiting patient outcomes and the health budget.

[1] We note NCI's data collections are comparable in scale to those of NASA's publicly-accessible collections.
[2] Demonstrated by the ability to align 1,200 human genomes for the Garvan Institute in an overnight run—a genuine "big data" task possible only with the use of a supercomputer and Australia's fastest filesystems.

- *Future Industries* – HPC is helping Australians to lead the world in applied R&D focused on: (i) increasing nutritional quality and crop yields by enhancing photosynthetic efficiency in grains; (ii) prototyping a range of renewable energy technologies; and (iii) *in silico* design and characterisation of "smart materials", with myriad technological applications.

As Australia's most highly-integrated e-infrastructure, and its most highly co-invested, NCI's commitment to an integrated national eResearch infrastructure, shaped by national research priorities, and deeply engaged with research communities and institutional partnerships, is manifest in the above portfolio. However, the integration which is envisaged in the Issues Paper, under the heading "Data for Research and Discoverability" (Ch. 11), is focused entirely on data-centric research. This is unnecessarily and unhelpfully narrow in its conceptualisation. As is stated in a recent report to the US National Science Foundation, "For many scientific disciplines, the issue is not whether to use data or simulation, but how the two will be used together". This is evident in the domain requirements described throughout the Issues Paper.

NCI therefore paints a vision of a more encompassing, more highly-integrated e-infrastructure, termed an integrated digital ecosystem. This is realisable by bringing together computational and data-intensive capabilities to establish a more strategic, more powerful infrastructure. In so doing, it addresses a capability gap in the Issues Paper, whereby the peak facilities and their big data capabilities were overlooked as critical elements of an infrastructure for data-intensive research.

An integrated digital ecosystem is required as the infrastructure foundation for the solution of the most challenging of research problems, and the means by which to support multi-, inter- and trans-disciplinary approaches to research throughout the sector—in universities, science agencies, medical research institutes, and industry.

The envisaged integrated digital ecosystem comprises:

- Platforms (peak, private and public cloud, and perhaps also institutional facilities), with strong infrastructure integration, that deliver computational resources of varying capability, and provide storage, of varying capacity and performance, for holding nationally significant and institutional data assets;
- Service delivery through the platforms/facilities, etc., for research that is data-dependent/intensive and reliant on analysis/analytics (in addition to data management, curation, provenance, interoperability, publication, virtual laboratories, etc.), or reliant on advanced modelling and simulation capabilities, or both;
- Policies and processes for ensuring data accessibility, reliability, quality, etc., together with the dissemination, proselytisation and accreditation of best-practice data management and curatorial standards.

A critical element in this ecosystem will be the National Advanced Computing Service (NACS), to evolve from the peak computing facilities (NCI and Pawsey) to provide Australia with an internationally-competitive peak computing infrastructure, best practice, converged computational and data services, and a critical mass of expertise.

More so than ever, skills and expertise are critical to leveraging research outcomes and maximum impact from the large infrastructure investments. This is particularly so for HPC where, today, supercomputer codes require extensive reworking to realise, in real world applications, a significant fraction of the latent performance that is available in the new technologies. Exemplifying this is work at NCI which enhanced 10-fold the scalability of an ocean modelling code, allowing it to run at far higher resolution, so enabling new and deeper understanding of the energy dynamics of the Southern Ocean, and its impact on the global climate system.

A much enhanced computational and data science capability (i.e., expertise) is required, comparable to that in national laboratories or major centres in the US, UK, Europe, China and Japan. Such hubs of expertise make substantial contributions to the research and innovation economies of these nations, enabling program-scale R&D in areas of national priority—many of which are rooted in big data and HPC technologies. Evolution towards a national capability will require institutions to invest together with the Commonwealth, in a capability of national scale and impact, in contrast to sub-scale local capabilities.

Co-investment, which augments and/or complements Commonwealth funding, will be an important element in delivering and tailoring an integrated digital ecosystem of the requisite capability. The ecosystem, therefore, will be a shared responsibility of government (through NCRIS) and institutions, with governance which incorporates institutional ("skin in the game") representation, and expert independent advice being required.

Resources to establish or expand service components, implement or evolve software capabilities (both computational- and data-centric) for domain-specific applications, provision appropriate storage for data (particularly as an input to research), etc., must be prioritised through research domain-focussed advisory structures. These would act to strongly align implementation with national research priorities, and with the priorities of NCRIS capabilities and leading research institutions – consistent with their roles as both stakeholders and co-investors. The future resourcing plan must be research-driven and holistic in its application, in contrast to past implementations whereby NCRIS funding for software and skills was focused largely on data-centric activities. In particular, investments in software must provide for both data and computational requirements, or a combination.

The role of the Commonwealth/NCRIS and institutions in supporting the components of the ecosystem must take into account the increasing commoditisation of tier 2 (non-peak) infrastructure, progressively allowing this to become an institutional responsibility with time, and migrating into commercial cloud infrastructure. This will maximise outcomes ensuring the efficient and most appropriate use of NCRIS monies. Infrastructure aggregation can further enhance robustness and efficiency, in turn maximising the resources available for research-facing services and user support.

Finally, funding certainty is critical to the maintenance of a world-class infrastructure, particularly for peak HPC systems, the competitive lifetimes of which are at most 3-4 years. The peak facilities presently depend on co-investment for the bulk of their operational costs and the funding to employ world-class expertise; certainty of an ongoing Commonwealth commitment (at an appropriate level) for refreshing capital infrastructure is a prerequisite for the retention of strong co-funding.

The vision outlined, and characterised in the diagram of Sec. 2.6, is ambitious, and is needed to provide the infrastructure foundation for the solution of research challenges in all fields, with the integration to support inter- and transdisciplinary approaches. Nevertheless, it is realisable by building on present strengths and advantages, reshaping the solid progress to date, and by combining Commonwealth (NCRIS) funding with co-funding from research organisations and NCRIS capabilities to establish a capability of genuine value, quality and innovation.

Our submission comprises an overview to characterise the integrated digital ecosystem, followed by responses to individual Issues Paper questions on matters of detail.

2. OVERVIEW: An Integrated eResearch Infrastructure Environment (Questions 30–35)

This section addresses the Questions 30–35 of the Issues Paper in a holistic manner. The need to do so arises out of the gulf between the conceptualisation expressed in the Issues Paper, and the integrated digital (computational and data) ecosystem advocated in this submission. Sub-headings below are cross-referenced against relevant questions in the Issues Paper.

2.1 Introduction (Q.30 and Q.33)

In recent decades, almost every discipline in science and engineering has been revolutionised by both large-scale simulation and modelling, and the capacity to collect and analyse unprecedented volumes of data generated by scientific instruments and sensor networks of increasing capacity and sophistication. Now, it is not a matter of whether to use modelling/simulation or data, but of how the two can be used together most effectively.

Today, the greatest research challenges, which focus on societal, environmental, and economic outcomes as well as the advancement of knowledge, increasingly require interdisciplinary and transdisciplinary approaches. The most difficult, and potentially the most impactful, research problems no longer respect discipline boundaries. They require convergent[3] approaches that integrate insights and methods from more than one field to create powerful frameworks for investigation and discovery.

The extent to which highly-integrated infrastructures are increasingly needed to support converged, or system, approaches to research is exemplified by the breadth and depth of NCI's engagement across the entire spectrum of "big science".

Today, the NCI environment brings together a supercomputer, a high-performance private cloud, Australia's fastest filesystems, and an expert team of approximately 60 staff, skilled in the technology, its operations, and deeply engaged with research communities through a team with expertise in high-performance computing and data innovation, collections management, data curation and interoperability, virtual laboratories, domain-specific data services, and visualisation. In the context of data, this has led to a well-integrated, standards-based, data services environment.

> NCI now supports over 500 projects involving 4,000 researchers from: 31 universities undertaking research through over 200 ARC- or NH&MRC-funded projects annually (including 6 Centres of Excellence)[4] and generating more than 500 refereed publications annually; eight NCRIS Capabilities; four national science agencies (including BoM, CSIRO and GA); a growing number of medical research institutes, and industry R&D users[5]. The national benefit accruing from the national investment in this support is manifold, and includes numerous industry outcomes, some of which have been highlighted in the summary. Cash co-investment, as a measure of value and impact of the service, is $12 million in 2016.

2.2 A more holistic view of integration (Q.30 and Q.33)

Increasing convergence of research requires an *integrated digital ecosystem* of computational and data infrastructure and services—in contrast to, but a generalisation of, the integrated research data

[3] US National Research Council, "Convergence: facilitating transdisciplinary integration of the life science, physical sciences, engineering and beyond", http://www.nap.edu/download/18722 (2014)

[4] In 2014, the ARC- and NH&MRC-funded projects supported by NCI represented approximately $56 million in funding support.

[5] NCRIS Capabilities include AAL, AMMRF, ANFF, APN, AuScope, BPA, IMOS, TERN. Science Agencies include the Bureau of Meteorology, CSIRO, Geoscience Australia and the Australian Antarctic Division; MRIs include the Garvan Institute, the Victor Chang Institute and the Australian Genomics Health Alliance (AGHA).

system advocated in the Issues paper (Sec. 11). The misconception that is inherent in ignoring or underplaying the nexus of data and advanced computational methods is a backward step – one that arises by viewing the research system through solely a "data lens", as is implicit in the ARDIS quotation "data is central to all research" [6] which prefaces Chapter 11. This overlooks the essence of the scientific method, central to which is the ability to formulate, test and modify hypotheses—with the capacity to theorise and model, and to make predictions *through modelling, simulation and data analysis*, being critical elements in contemporary science.

The misconception is compounded by erroneous references in Chapter 11 to NeCTAR, RDS and ANDS as being *the* totality of the existing data infrastructure investments. When considering e-infrastructure of national scale and significance, the peak computing investments[7] (NCI and Pawsey) are central features of the national research data infrastructure landscape, with each having made key contributions to data science in their domains of specialisation, and to Australia's prominent position in global data science.

In the broader context of data, it is convenient to regard national growth and mainstreaming of "good" data management practices as falling into three, quasi-independent streams: (a) the ongoing evolution of data management in domains independent of infrastructure programs; (b) the development of data policy and standards fostered by infrastructure programs, (c) the development of data management driven by active research methods embedded in e-infrastructure platforms. All three streams have made substantial contributions to a rapidly evolving national data scene. Contributions of a peak facility like NCI fit squarely in stream (c), and driven largely by curatorial and accessibility aspects of big data in an active and interoperable environment, are highly valued by the research communities that are serviced. The Roadmap will need to consider the dynamics of all three streams when thinking through the role of infrastructure capabilities and their funding needs.

To conclude, while the goal of a "more integrated, coherent and reliable system" which is "informed by, utilise[s], integrates and build[s] on existing investments" (Sec. 11), is strongly supported, the context must to be broadened to establish a *well-integrated digital ecosystem*, rather than solely an integrated data system.

2.3 An Integrated Digital Ecosystem (Q.30 and Q.33)

The integrated national digital ecosystem for research comprises:

A. Robust infrastructure platforms within which computing and storage infrastructure are well-integrated, and overlaid with the skills and expertise needed to establish the service suite to support computational and data-dependent/data-driven research; and
B. A framework to ensure expert curatorial services for data management and provenance needed to ensure that data assets (both national and international) are findable, accessible, interoperable and reusable, in keeping with their value to research communities.

These perform separate and necessary functions within a digital ecosystem:

A. The platforms would be implemented as facilities that provide both tier 1 (for high-performance methodologies needed to meet high intensity computational and data

---

[6] Research Data Infrastructure Committee, *Australian Research Data Infrastructure Strategy* (ARDIS), Australian Government Department of Education (2013).

[7] The Super Science Initiative Funding Agreements for each of NCI and Pawsey included clearly defined, nationally significant data roles; both were later referred to as part of the national research data infrastructure in the ARDIS (2013), which is quoted from in the Issues Paper. Each holds Australia's largest collections of openly accessible research data in their domains of specialisation, comprising data generated within, or processed, by them and overlaid by a comprehensive range of data services.

requirements) and tier 2 (cloud-based platforms for less intense requirements) computational and storage infrastructure, overlaid with interoperable data services – all of which is enabled by the expertise to implement the suite of software services and virtual laboratories that is required; and

B. The data management framework picks up core data discoverability and curatorial standards functions, which have been funded to date through ANDS.

This structure accords with the conclusions of the eResearch Framework (eRF) Project which put forward, as major components:

- National Computational Facilities (NCF): "construct[ing] powerful digital resources", "build[ing] on computational modelling, visualisation and data analysis", being "a data-generating and data-using infrastructure" and "implemented as large facilities"; and
- Australian Research Data System (ARDS): "improv[ing] accessibility, value and use of quality research data", "support[ing] co-operative possession of digital data assets, empowering their use and re-use", being "a data-organising and data-using infrastructure supporting an Australian research data cloud" and "implemented as a co-operative system"

Each would have the capacity to "hold research reference data and publish Australian research data", with "highly compute-intensive data holding being migrated into the NCF".

2.4 Globally-recognised model for converged services (Q.30 and Q.33)

It is important that the 2016 Roadmap recognises that the integration of infrastructure and services, envisaged in the Issues Paper, is already exemplified in Australia's peak computing facilities. This is most notably demonstrated by NCI which provides Australia's most highly-integrated e-infrastructure environment[8]. It is from both the NCI model as it is today, and the interrelationship of the NCF and the ARDS, outlined in the eRF project, that the conceptualisation of an integrated digital ecosystem has evolved.

- Such an approach is the central recommendation in the recent *Future Directions* report[9], from the US National Academies.
- The report articulates this as current world's best practice, which takes the form of converged computational/data services. This *fusion of big compute, big data, and expertise* is now referred to as *Advanced Computing*.
- Such directions are being pursued by leading US peak facilities, including those funded by the National Science Foundation and those operated by the Department of Energy, and elsewhere in the world.
- NCI's research- and mission-driven evolution over the past five years places it in the vanguard of Advanced Computing.

Despite such directions being not being taken into account in the Issues Paper, it is vital that the next generation of Australia's e-infrastructure be configured with such attributes in order to provide a platform that will sustain and advance the competitiveness of Australian research into the future.

---

[8] The NCI environment comprises a tightly-integrated platform (supercomputer, a high-performance private cloud for data-intensive services, fast filesystems), a comprehensive service portfolio in simulation and data services (including data curation and management, domain specific services, virtual environments and visualisation), and a large volume (10+ petabytes) of national and international reference collections—all of which are curated, accessible through catalogues, and significantly interoperable and reusable.

[9] *Future Directions for NSF Advanced Computing Infrastructure to support US Science and Engineering in 2017-2020*, National Academies Press, http://www.nap.edu/21886 (2016).

2.5 A National Advanced Computing Service (Q.30 and Q.33)

The realisation of an integrated digital ecosystem will require the peak computing facilities (i.e., the NCF of the eResearch Framework) to evolve to form a coherent *national advanced computing service* (NACS). In addition to its role as the provider of tier 1 national high-performance research computing services, the NACS will form a critical service platform of a national research data system.

While the evolution of governance to support the NACS may take some time to evolve, the need to replace ageing infrastructure—particularly NCI's supercomputer—is especially urgent, and should

> <u>A National Advanced Computing Service: ***Vision***</u>
>
> A future *National Advanced Computing* Service, evolved from NCI and the Pawsey Centre, is needed as a key component of the integrated digital system to provide infrastructure of a scale and performance, and embody high levels of expertise, to:
>
> - Provide Australia with an *internationally competitive, high-performance compute/data production service*, the combined capability of which will rank in the world's top 20 facilities;
> - Offer *best-practice, converged/integrated computational and data services* (including curated national collections), shaped to meet the needs of Australian research and innovation, facilitating collaboration around large and complex data, and made available through both merit and priority access mechanisms;
> - Sustain *a critical mass of expertise* that is respected nationally and internationally for the *quality and the innovation* of both the *operational services* and infrastructure integration; the *research-facing service portfolio* scoped to deliver transformative outcomes; and excellence of its contributions in computational and data science;
> - Have the *remit to represent Australia in international peak computing and data collaborations/forums, e.g.,* www.exascale.org/bdec (big data and extreme-scale computing)
>
> Its mission, to provide:
>
> *World-class, high-end, advanced computing and data services for Australian research and innovation*, will be shaped by national research priorities, the priorities of NCRIS capabilities, and those of its institutional partners.

not be jeopardised by governance considerations in the short term.

2.6 Elements of an Implementation Model

Research-directed prioritisation, Advisory structures, and Resource decision-making. (Q.30 and Q.33)

To function effectively, the directions taken by of the NACS and the tier 2 platform and service components of the ARDS must align with national priorities, NCRIS and research community priorities, and institutional priorities, and also reflect current and potential co-funding and co-investment dependencies.

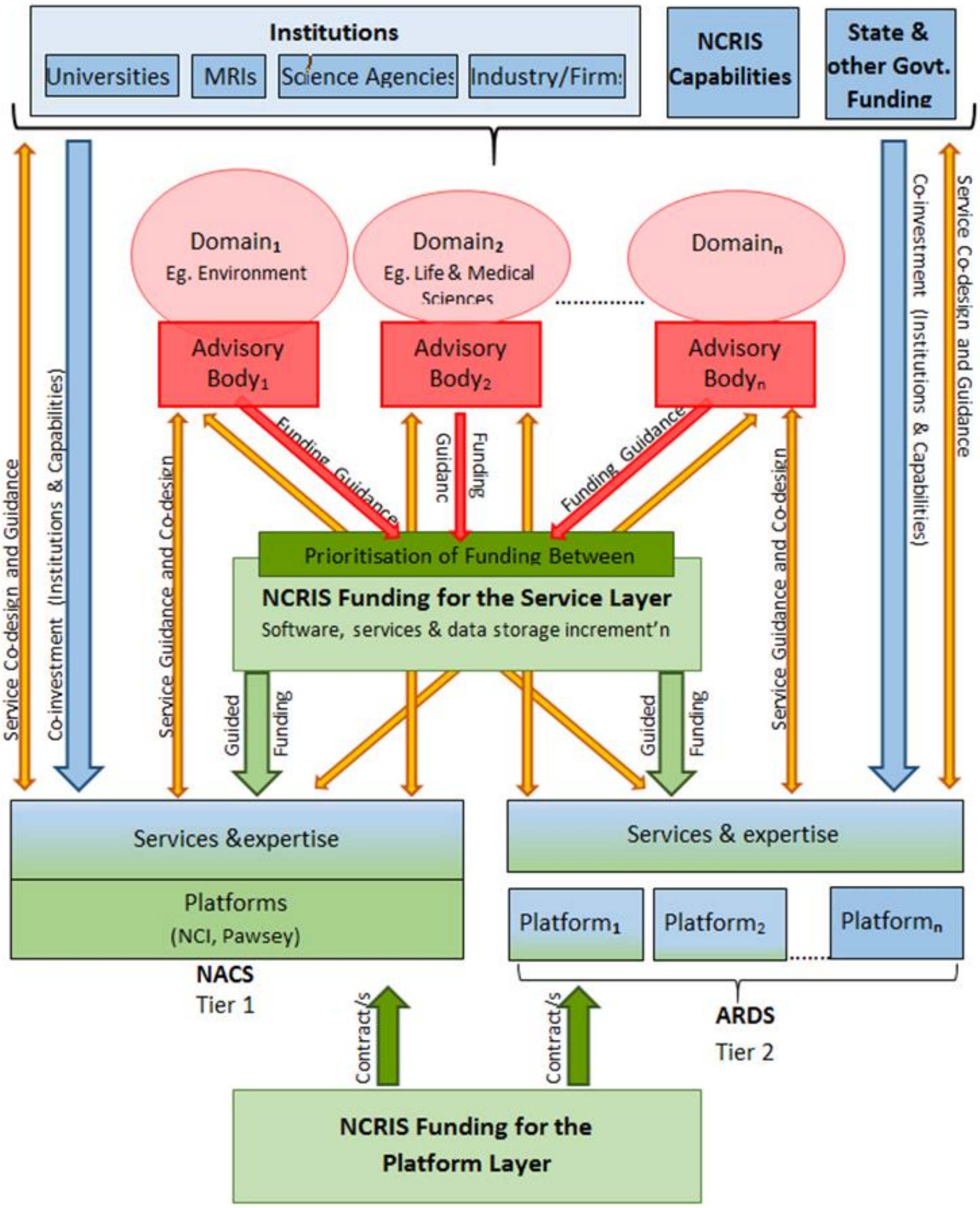
Research-directed prioritisation is already a feature of the NACS components, and this will benefit the ARDS components by setting goals that align with those of NCRIS and other national priorities. This will simultaneously drive out the fragmentation inherent in service delivery through disjoint components (presently, NeCTAR, RDS and ANDS).

The diverse and complex requirements of research communities will require service contributions from each of NACS and ARDS components—in the form of tier 1 computation and/or storage and data services from the NACS, and tier 2 storage and data services from the ARDS.

It will be necessary that the NACS and ARDS components interface with the communities they serve *in a harmonised manner*, strongly informed by domain-specific advisory structures, which they should establish, and which bring together the needs of national research priorities, NCRIS capabilities and research institutions which are substantively associated with each major research community.

These advisory structures will:

- Ensure fitness-for-purpose in the design of the service portfolio;
- Play a leading role in prioritising resource requirements which are essential inputs to research in each domain, e.g., software and service development, and recommend funding allocations for data storage, domain-specific user support, etc.



This avoids flaws of previous implementations which were often fragmented and improperly aligned with national and NCRIS priorities. The advisory structures, to be set up as joint responsibilities of the domains and platform/service providers, would provide their advice and prioritised requirements in alignment with, and on the same timeframes as, NCRIS Project Reports and Annual Business Plans.

Co-investment (both institutional, and from NCRIS capabilities) must inevitably play a significant role in the prioritisation of resource allocations and augmentation, with the Roadmap needing to refer to the manner in which co-investment and Commonwealth funding can be blended.

Guidelines that preclude institutional cost-shifting will need to be considered.

The diagram on the previous page outlines at a high level a realisable national *integrated digital ecosystem*, elements of which are discussed throughout this section. Details regarding the specifics of the implementation model are provided in **Attachment 1**.

Holistic purposing of research support funding (Q.30 and Q.33)

An integrated digital ecosystem for research requires that distinct but interrelated components be brought together—of which well-integrated infrastructure platforms are but one. Software, data, and user support, in a range of forms, are also critical components of the ecosystem, shaping the service portfolio to meet the generic and particular needs of each research community.

Unfortunately, the manner in which the Issues Paper is structured results in high-performance computing being referred to solely as a platform (in Chapter 10: Underpinning Infrastructure), whereas Chapter 11, which describes the national data system, refers to the full spectrum of services (software services, virtual laboratories, etc.) and the resources needed to provision these.

The risk of perpetuating the problems which exist today—whereby NCRIS funding for software and user support is partitioned into the NeCTAR and RDS projects, thus localising its use in data-dependent activities (e.g., VLs, domain data services)—must be avoided.

Therefore, a holistic approach to software investments is required—one that is able to support all classes of research software (i.e., both data-centric and computational methodologies) necessary for priority national research—including, for example, the national weather/climate modelling suite ACCESS, which is referred to in the Environment section of the Issues Paper[10].

This holistic approach is also needed for domain-specific user support in both data and computational capabilities. An example of the latter might be in the domain of advanced materials, where computational support might be provided through a combination of the budgets of relevant NCRIS capabilities (i.e., ANFF and AMMRF), a contribution from an eResearch line, and institutional co-investment.

A funding pool which can be used holistically to enhance services, software support and user support, together with provisioning collections storage, is needed—along with advice on its prioritisation being through the advisory structures mentioned previously. Co-investment, as a proxy for value, will be an important signal in prioritising the use of this funding pool.

[10] ACCESS provides the foundation of national research in climate system science, and also generates Australia's data contributions to the international climate model comparison project CMIP, the totality of which underpins IPCC assessments on global climate change.

ARDS as a "cooperative system" (Q.30 and Q.33)

The ARDS was conceived as a "cooperative system" in the eResearch Framework (eRF), with the platform and data services commissioned from tier 1 and tier 2 facilities. The need to operate as a *virtual provider* of services, while offering the same accountabilities as a facility-based operator, will be a challenge. The changes required to evolve the infrastructure and service delivery framework, currently provided by the various "nodes" of NeCTAR and RDS are likely to be substantial.

- Such changes will require independent, regional eResearch operators, together with the data fabric of universities and perhaps other institutions, to coalesce into a national system.
- The need to balance necessary improvements in operational efficiency, which is effectively achieved by aggregation (as commented on by the eRF), against a likely need to have a point-of-presence in each state might best be handled by *concentrating infrastructure* but appropriately *distributing expertise and service delivery*. Dilution and unproductive duplication of effort might be avoided by designating domain centres of expertise in specific geographic locations.
- It is possible that much, if not all, of the tier 2 infrastructure provision for ARDS may be able to migrate into commercial cloud providers in future years, largely simplifying the attainment of these goals.

Security and Service robustness (Q.32 and Q.35)

<u>Security</u>

The high and growing dependence of research on national e-infrastructure raises its profile as an attack target—both physical and cyber.

Increased physical security of data centres, particularly those from which the peak systems operate, will need to be considered.

Facilities (both tier 1 and tier 2) will need to work with the Australian Cybersecurity Task Force (Australian Signals Directorate) and other agencies of the Australian Government to conform with provisions of the Government Information Security Manual, and then demonstrate conformance of operational practices with these provisions. Procedures to secure, and test the efficacy of, large datasets and large (multi-petabyte) filesystems, and the transfer of large data volumes, particularly through government agency gateways, will be required.

Conformance with policies for minimum standards of data security is already a requirement in some domains – notably for data of human origin. In this case, certified adoption of the US HIPAA standard will be required to facilitate international collaboration.

<u>Infrastructure robustness and reliability</u>

Ensuring that national e-infrastructures have robust, reliable operations is essential if researchers and co-investing organisations are to rely on them. Simultaneously, this is a pre-requisite for ongoing co-investment, and therefore essential for the sustainability of facilities.

Increasing the robustness of the data centres from which peak facilities operate may require retrofitting to increase the levels of redundancy/duplication of power, cooling, networking etc. to increase fault tolerance. For the facilities of ARDS, a progressive migration of infrastructure into commercial data centres, or an increasing use of commercial cloud infrastructure, is the way forward.

Infrastructure aggregation enables the establishment and retention of a critical mass of staff with the diversity and depth of expertise to ensure production-grade operations. Services that require

24x7 levels of rigour may necessitate: (a) staff being "on call", (b) upgrades from "next business day" to higher frequency provision of vendor service, or (c) the inclusion of some aspects of operations from higher-tier installations such as a commercial cloud. Each of these this will increase the cost of operations, but will become a necessity if the e-infrastructure is to support a research and innovation system in which data services are expected to be "always on", and in which industry and government agencies are significant users and beneficiaries.

Increasing dependency on the national e-infrastructure will bring with it also the need for increased infrastructure redundancy, and some duplication in order to support disaster recovery functions, in the event of a catastrophic failure or event. Within the peak facilities, the presence of two national facilities provides a way of mitigating risks associated with the total failure of one, and also affords an opportunity for each to act as a data backup for the other. Consideration of a modest level of infrastructure duplication is needed to provide for disaster recovery. For non-peak provision, cloud environments implicitly provide this robustness and the ability to replicate data and re-deploy services.

Tier 2 considerations (Q.30 and Q.33)

For the research computing platforms, while there remains market failure for the Tier 1 component, the same cannot be said for tier 2. Infrastructure costs for a tier 2 capability are within the capacity of institutional budgets, or are available as a competitive service from commercial cloud providers. NCI suggests an exception might be made for specialised, tier 2 facilities that are aligned with a major instrument which has a close association with an NCRIS capability.

For Commonwealth funds to continue be invested in tier2 data infrastructure, either efficiency through aggregation of current infrastructure and activities must be achieved, or significant additional co-funding from state governments and institutions will be required to reduce the call on Commonwealth funds.

Commonwealth funds might be used effectively to transform tier 2 infrastructure provision into a commercial cloud environment, in line with the increasing cost effectiveness of this mode of provision, and also policies within government.

2.7 Skills and Expertise: an enhanced, national computational and data sciences capability (Q.32 and Q.35)

Skills and expertise are critical to maximising impact from the large investments in the NACS infrastructure. This is particularly so in the HPC environment in which:

a) There is no longer an omnibus processor architecture for which software can be developed or optimised; and
b) There is a much greater need than previously, to rework and optimise supercomputer codes in order to realise the latent performance that is inherent in the new technologies.

Exemplifying this is recent work at NCI which has enhanced the scalability of an ocean modelling code from the use of 960 processor cores in parallel to more than 10,000. This makes possible far higher-resolution simulations, thereby enabling new science, and a greater understanding of the energy dynamics of the Southern Ocean, and its influence on the global climate system. Comparable gains in outcomes can be cited from work to enhance the interoperability of large data collections in the arena of data-intensive science.

A significant increase in the skills base is needed to reduce the gap between infrastructure capability and researcher ambition. The importance of holistic approaches in prioritising resources for expert support must be re-emphasised.

At present, the national skills base in advanced computing is sub-scale, dispersed and lacks focus. While there are staff with excellent skills, the numbers are insufficient to attempt other than research support at the project-scale. We advocate that the skills base must increase to a level whereby the requirements of program-scale research and innovation, in areas of national priorities, can be addressed. Exemplifying this might be support for the ongoing development and performance optimisation of the national climate and earth system science software suite, ACCESS, or the implementation of national repositories and analytics services for human genomics, and pipelines for processing and serving environmental, geoscience, and earth observation data, amongst others.

This requires additional staff who are research-trained, possess expertise in computational and data methods, and bring experience in an application domain. The emphasis is on *research enablement through exploitation of advanced computing*. The purpose *is not* research into advanced computing, or in the domains that it supports.

Our vision is for a national hub of computational and data science providing a critical mass of expertise through which to:

- Drive solutions to national R&D challenges to advance Australian research and economic competitiveness;
- Enhance the Australia's participation in international collaborations in advanced computing and its applications (e.g., the exascale.org project, collaborations with international science agencies such as NASA, NOAA, LLNL, NSF, ESA, etc.); and
- Play a major role in national training and education for advanced computing— to enhance research outcomes and competitiveness, and assist in skilling a future workforce for jobs in an increasingly information/data-rich world.

Critical masses of skills in advanced computing are found within peer international facilities and the national laboratories in advanced economies— notably the US, EU, UK, China and Japan, and have been in place for many years. Such hubs make substantial contributions to the research and innovation economies of these nations, enabling program-scale R&D in areas of national priority— many of which are rooted in big data and HPC technologies. Evolution towards a national capability will require research institutions to invest with the Commonwealth in a capability of national scale and impact, in contrast to sub-scale local capabilities. It is important that the Roadmap send signals to this effect if Australian R&D is to benefit from an expert computational/data science capability of genuine national significance and value.

3. INPUT IN RESPONSE TO 2016 ROADMAP ISSUES PAPER QUESTIONS

The input in this section responds to the material in the 2016 Roadmap Issues Paper, and is provided as responses to selected questions (1–14).

Q. 1: Are there other capability areas that should be considered?

The focus on the areas of health and medical research appears to have omitted other important areas of the life sciences, including plant science and agriculture.

Q. 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national infrastructure?

The listed characteristics are appropriate but omit the valuable influence of co-investment in strengthening governance through "skin in the game". In cases where high levels of co-investment are needed to sustain the operations of a facility, the governance needs to reflect the shared responsibility between government and stakeholder partners which have an interest in the successful carriage of the project, and which align it with national research priorities and NCRIS capabilities. Healthy governance thus requires a balance of expertise and skills, including from both stakeholders and independent members.

Into the future, the eResearch capabilities, particularly those that are derivative of the current ANDS, NeCTAR and RDS projects, will need far greater alignment with national research priorities and NCRIS capabilities, than has been the case to date. This proposal advocates advisory structures which guide resource allocations in alignment with priorities.

Further, significant co-investment enhances the rigour of governance, and hence the quality of outcomes—something which been lacking for a number of the current eResearch investments.

Q. 3: Should national research infrastructure investment assist with access to international facilities?

Q. 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

Combined Response

In general terms, the answer is indisputably yes, thereby providing access to facilities that may not exist in Australia, or be of the same capability.

In the realm of advanced computing, in the experience of NCI, there are no mechanisms by which a nation can buy into the HPC infrastructure of another. Previous discussions with PRACE of Europe reveal that the European perspective is that the provision of advanced computing infrastructure is a sovereign responsibility. This situation may evolve in coming years, with shared international (exascale) facilities in particular research domains.

While international repositories have been established in some domains, growing data volumes are beginning to overwhelm the original design concepts and funding base—leading to distributed or federated international repositories. Despite the speed of modern transcontinental networks, their bandwidth is but a tiny fraction of that available within a data centre, and so analytics and deep learning from massive collections will continue to require the use of large-scale (on-shore) facilities which exploit in-machine-room bandwidth.

Q. 5: Should workforce skills be considered a research infrastructure issue?

Q. 6: How can national research infrastructure assist in training and skills development?

Q. 7: What responsibility should research institutions have in supporting the development of research infrastructure ready researchers and technical specialists?

Combined Response

In the eResearch domain, the skills required to enable research are precisely those required in current and emerging jobs in an information/data-rich world. The requirements of the current workforce, let alone a future workforce, are outstripping the rate at which education and training can generate the requisite skills. Competition for skills is fierce, with industry competing with government agencies, which in turn are competing with universities and research infrastructure providers.

That said, the research training system, in tandem with the research infrastructure facilities, has the capacity to increase the rate of skills generation. With graduate students and postdoctoral fellows constituting the bulk of hands-on users of national research infrastructure, strengthened training and the introduction of formal courses will have beneficial flow-on effects for the workforce. The UK, for example, now requires that research students receive training in eScience methods, providing them with valuable skills to exploit during their research, but also providing the beginnings of alternative career pathways after graduation.

As a national capability, NCI provides general, specialised and domain-specific training throughout Australia. While very willing to increase the volume and diversity of training, NCI finds itself limited in its capacity to do by funding limitations—the ongoing consequences of an earlier funding regime which requires it to raise the bulk of its operational costs via co-investment, thus limiting resources which can be devoted to research-facing services, including training. Restructuring the allocation framework for NCRIS recurrent funding to resolve this disadvantage will alleviate this problem.

We note also that the value and importance of the specialised “para-research” skills of staff working in research infrastructure support roles is insufficiently recognised. Such roles are integral to environments which provide the crucial underpinnings of research and innovation, and yet, in an absence of financial certainty, the working conditions of such staff are comparably uncertain—equivalent to those of staff supported by research grants. Issues of career recognition, training and progression need consideration as a critical component of retaining and building a research infrastructure capability.

In the case of the e-infrastructure, the focus of this submission, aggregation of staff, where possible, into larger units or facilities would facilitate career development pathways.

Q. 8: What principles should be applied for access to national research infrastructure, and are there situations where these should not apply?

Access measures at a facility like NCI are shaped by principles of research excellence (merit) and research priority where:

- Merit refers to the absolute excellence of the research (aligned with processes of the national research councils), and is complemented by consideration of the suitability of the infrastructure to support the research project;
- Priority refers to the absolute significance of the R&D in national or global terms.

Merit-based access is typically used for projects originating in universities (often supported by ARC and/or NH&MRC grants). Priority access largely reflects the needs of government science agencies, for which merit-based access is an inappropriate and inherently risky mechanism for supporting their program-scale R&D.

Priority allocations would appear to be a preferable means by which to support the work of NCRIS capabilities, in which merit-based decision-making is either inappropriate or contains an element of risk. In this situation, an NCRIS capability might acquire a share of resources (computation, storage, etc.) and provision access via its own processes, or establish a community resource, e.g., a bioinformatics pipeline. In present circumstances, where co-funding is required to meet the bulk of the recurrent costs (discussed below), an access charge, equivalent to that of the subscriptions for NCI partner shares, would need to be levied to support such access.

The current access model at NCI has been shaped by the EIF funding environment under which current peak facility was established. From the outset, co-investment was required to provide for the totality of operational funds, although mitigated recently by NCRIS funding which now sees the recurrent costs borne in the ratio 2:1 — ~$12 million p.a. in co-investment, to ~$5.5 million from NCRIS.

The need to meet the high operational costs of a peak facility has shaped a subscription-driven, shares-based access model in which:

- 80% of the access is provided to subscribing partners,
- 20% share is contributed *pro bono* to merit-based allocations, largely through the National Computational Merit Allocation Scheme (NCMAS), which provides the "computational" complement to the granting schemes of the national research councils.

A modest amount of commercial access for industry R&D is provided through commercial service contracts with full cost recovery, but with concessionary provisions for start-ups.

Access to advanced computing in Australia, through NCI, is somewhat different to that found elsewhere in the world. Indeed, to some extent, the role served by NCI is a hybrid of what is undertaken in the USA by the National Science Foundation facilities (through XSEDE) and the facilities operated by the national laboratories of the Department of Energy.

While the present arrangements work reasonably well, the absence of a clearly defined funding line to support the 20% merit-based access share, available to researchers at no charge, remains problematic.

- Firstly, this "free" share is subsidised by largely by partner subscriptions, and to a lesser extent by NCRIS, and by the ARC (through a NCI-led LIEF Grant). This partner subsidisation can be a source of friction at times to those organisations paying full subscriptions.
- Secondly, the present funding model is an impediment to being able to increase the merit access share. The absence of a designated funding line to support merit-based access is at the heart of this.

Accordingly, present arrangements are in need of reworking. Elsewhere in the world, national research councils (e.g., RCUK in the UK, NSF in the USA) are strongly involved in allocating and funding merit-based access as a supplement to research grants, and progress in this direction, or some equivalent via NCRIS, would be desirable.

Q. 9: What should the criteria and funding arrangements for defunding or decommissioning look like?

As research priorities change, so will priorities for ongoing infrastructure or its continuation.

In line with technology developments and market forces, some infrastructures previously funded by government may transfer to commercial provision as a commodity which can be dealt with as an institutional expense. The rise of commercial cloud computing is such an example.

Long-term planning for ongoing funding, or defunding, should be undertaken through regular reviews. In the case of advanced computing infrastructure, the natural investment cycle is either 4 years, or 6 years, inclusive of a mid-term upgrade at the 3-year mark.

A review undertaken 18-24 months before the end of the cycle will provide an opportunity to decide on the refunding of the infrastructure, or the winding up the infrastructure, with sufficient time to do so in an orderly manner. Commonwealth funding via NCRIS may be required to assist with the winding up of facilities, e.g., paying out staff entitlements. This comment is particularly pertinent to those facilities which are substantially reliant on co-funding to support their operations.

Q. 10: What financing models should the government consider to support investment in national research infrastructure?

Funding certainty

Both certainty of funding, and its frequency, are needed to maintain an internationally competitive infrastructure, and the base of skills and expertise needed to leverage research outcomes from the infrastructure.

Certainty is equally critical to the ability to sustain and develop co-investment, which for the peak computing facilities is a critical component of their recurrent operations, and particularly their expertise capability. In the absence of certainty, and especially in an environment in which the competitiveness of a facility is waning, co-investment will inevitably wilt as partner organisations seek better and more certain alternatives.

Co-investment

Unremitting demands from the research community for research infrastructure of increasing power and performance must overwhelm the capacity of public purse to provide for it. Increasing prioritisation will be required, as will an increasing emphasis on co-investment, as a proxy for value and as a means of contributing funding to either the recurrent or the capital costs.

In the realm of e-infrastructure, access to computing and storage is now increasingly a “commodity” at the second tier—which, in previous generations, would have been an infrastructure funded by government, but which today can be transferred to become an institutional responsibility. The rise of major commercial cloud computing providers allows recurrent funding to access capital infrastructure in increasingly cost effective ways at the commodity level.

At or approaching landmark scale, e.g., the procurement of a supercomputer, some appropriate form of financial engineering may assist government in flattening its cash flow, albeit with the necessity for an ongoing financial commitment.

Q. 11: When should capabilities be expected to address standard and accreditation requirements?

The application of standards increasingly provides the underpinnings for trusted, cross-disciplinary, interoperable research capabilities, as well as the secondary re-use of research outputs. These are fundamental to transparent and long-term reproducibility. As such, the adoption and progression of standards should be used, and enhanced, by capabilities, wherever possible. Better engineering of the entire workflow, in contrast to retrospective capture, is need to ensure born-connected implementation, through the close cooperation between the infrastructure capabilities and scientific domains. The adoption of quality assurance processes are also increasingly needed – to ensure that all elements of the integrated digital systems are fit-for-purpose for the application domain, and flexible for future increasingly connected, internationally linked and cross-cutting capabilities.

There is an increased need for trusted protocols to capture and make available provenance-based information through new standards. For example, there is an increased requirement for digital data repositories to be formally accredited using protocols such as those developed by Research Data Alliance, ICSU World Data System, and ISO. This will require further attention for data generation, data analysis workflows, and data services – including formal assurance of the management and curation of the underlying data. Such accredited digital repositories will provide the quality underpinnings for future publications, and increase the trustworthiness and reproducibility of Australian research.

Accreditation or certification of software, through NATA for example (e.g., bioinformatics processing pipelines), are an increasing requirement.

Enhanced information security practices, demonstrating compliance with the Australian Information Security Manual through regular security audits, will need to be a priority for the national research e-infrastructure in a world of increasing cyber vulnerability.

Q. 12: Are there international or global models that represent best practice for national research infrastructure that should be considered?

We limit our remarks to the field of advanced computing in which the deep integration of high-performance computing, big data and expertise is today widely regarded as a pre-requisite for extending the boundaries of inquiry, supporting multi-, inter- and trans-disciplinary approaches to research, and advancing research impact in every scientific discipline.

Such directions are advocated in various reports (referred to earlier in this submission) to national and pan-national governments, and which align with those which have been pursued by NCI for 4–5 years. Disappointingly, in the Issues Paper (Secs 10, 11) these important directions, which *must shape* our future e-infrastructure, have not been considered.

Q. 13: In considering whole of life investment, including decommissioning or defunding for national research infrastructure, are there examples, either domestic or international, that should be examined?

See the response to Question 9.

Q. 14: Are there alternative financing options, including international models that the Government should consider to support investment in national research infrastructure.

See the response to Question 10.

Schematic of a national eResearch Infrastructure System

The diagram below and the text following it, outline at a high level a realisable national *integrated digital ecosystem*, featuring the major infrastructure elements considered in section 2 of the NCI submission on the 2016 Roadmap *Issues Paper*. The integrated digital ecosystem is informed primarily by, firstly, NCI's experience as a heavily co-invested provider of converged high-end services for computationally- and/or data-intensive research, and secondly, the findings contained in the recent eResearch Framework.

The system represented reflects the following operational realities which apply to e-infrastructure if it is to support leading-edge science in delivering innovation outcomes. These realities have been made clear by a decade of NCRIS experience, including negotiating the transition to funding arrangements under the Super Science Initiative.

- Financially, operating national scale e-infrastructure requires a national scale of commitment—in which the shared responsibility of government and institutions is a critical element.
- Top-down decision making is required to ensure e-infrastructures provide services
  - Which align with national, NCRIS and institutional priorities; and
  - Are responsive to research community needs.

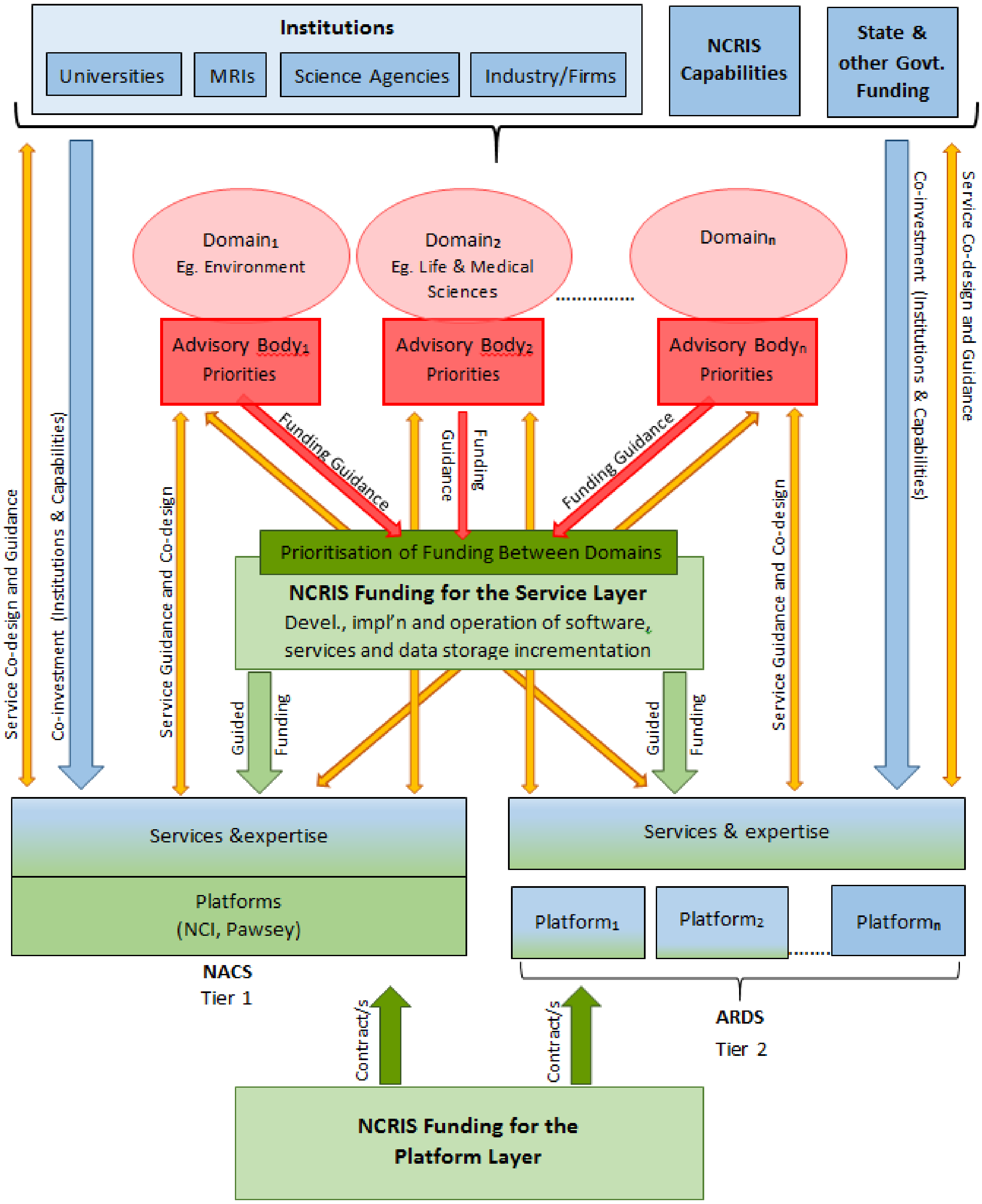
  This necessitates that national eResearch infrastructure be driven not by technological potentialities, but rather by the prioritised needs of researchers as expressed by bodies appropriately representative of communities.
- Research communities will require service contributions from both the NACS (Tier 1) and the ARDS (nationally integrated Tier 2) components of the system.

Responding to these realities requires operational structures to which national eResearch infrastructure investments will need to conform, if they are to effectively fulfil their mission in a manner that is relevant, aligned with priorities, and efficient.

A key feature of these operational structures derives from the nexus which exists between the provision of infrastructure services, and its attractiveness as a site of co-investment. Past planning and delivery mechanisms, via intermediaries, has led to some sub-optimal outcomes. Co-investment is a strong proxy for value, and can be an effective means by which to shape the resource allocations which determine the services.

The implementation model detailed below aims to lead to outcomes which ensure:

- A holistic approach, in which appropriate support is provided for *all* elements necessary to deliver an e-infrastructure capability.
- Ineffective allocations of resources are not made, ensuring funded infrastructure has the capacity to meet national requirements.
- e-Infrastructure in receipt of national funding provides services, shaped by research communities, and aligned with their priorities.



The diagram is colour-coded, with colours for decision-making and guidance mechanisms, and colours representing the flow of funding. Reading the diagram requires familiarity with its two key elements, detailed below.

Element One: Flows of funding

Funding comes from two main sources, to enable the operation of the underpinning e-infrastructure which both provides converged services in support of research, and comprises the backbone of a national data system.

- Green represents direct national funding – support from the Commonwealth, for the specific purpose of operating national-scale research infrastructure.

- Blue represents co-funding – real co-investment (often, although not always public funds) from the organisational budgets of parties other than the Commonwealth (i.e., via NCRIS). The co-investors of primary significance are universities, science agencies, MRIs, NCRIS Capabilities, and in the future industry.

Thus, the infrastructure platform for Tier 1 Advanced Computing facilities—for which there is demonstrable market failure and which together comprise the NACS—are funded by the Commonwealth through NCRIS. The platform component for Tier 2 Advanced Computing facilities, for which the market now increasingly provides, is able to be provided increasingly by institutions (or consortia), including as a service from commercial providers. National funding for this layer may be expected to diminish over time.

However, as has been detailed, computational and data infrastructure services arise from an integrated ecosystem, which includes a layer of expert services provided to support researchers to best exploit the computational and data platforms. In line with the operational parameters outlined above, this service layer is funded partly by co-investing bodies, who direct their own resources to it on the basis of their needs. In this system, co-investing bodies and other, agreed stakeholders act separately and in concert to direct funds to the NACS and/or ARDS service layer by two means:

- Bodies requiring a particular service co-invest directly in the NACS and/or ARDS as required, to access the service beyond what may be available freely—as is already the case, particularly for the peak computing facilities.
- Eligible stakeholders in a range of broadly defined research domains, contribute to the formulation of expert guidance. This guidance determines priorities for the allocation of national (NCRIS) funding for the service component of the NACS/ARDS.

So, to take the example of high-end support for climate and earth systems science, which is a national priority area, and for which NCI provides specialised support:

- The collective decision of domain stakeholders about priority requirements is a foundational input for the subsequent allocation of national funding to support the development and delivery of services which meet those requirements (including computational software and data services).
- This funding combines with, and may be influences by the existence of, co-investment from some or all of those same stakeholders, which include science agencies, universities and NCRIS capabilities. As is already the case, this co-investment in the NACS and/or ARDS may be for a variety of reasons, including ensuring unique or extra requirements are met, or augmenting an extant service so that it can be expanded to support their needs.
- The streams of national funding and co-investment combine to support the delivery of services of appropriate scale, quality, reliability, and innovation.

Element Two: Guidance mechanisms

For this system to be effected, a guidance mechanism is required. This is represented by the red features in the diagram, and should be established by NACS and ARDS jointly with the domains (research communities), helping to ensure fitness-for-purpose and harmonisation of service delivery.

- The light red ovals represent the established, identifiable views of the broader community of stakeholders in a research domain – for example, environment, medical and life science, etc.
- The darker red squares over the top of the ovals, comprise advisory bodies, whose guidance prioritises the requirements of the community, and informs the allocation of national funds to the service layers of the NACS and/or ARDS.

While these bodies represent broad research domains, it is important to note that some such bodies already exist. Astronomy Australia Limited, for example, is a mechanism that has effectively enabled coordinated decisions and action by the group of research organisations which are most heavily invested in astronomy research.

For the guidance bodies to strike the correct balance between ensuring funds meet organisational and domain-driven missions on the one hand, and are directed towards leading-edge science which most aligns with national priorities on the other, a balance of institutional representation and independent scientific expertise is required. A balance of representation would need to take into account alignment with strategic national priorities, research prioritised on the basis of excellence (e.g., ARC/NH&MRC Centres, Projects, Fellowships, etc.), and alignment with institutional missions (including science agencies). This can be achieved by recognising the alignment of different types of bodies and institutions, with the different priority types.

A non-infrastructure body supporting data curation practices

Although not represented in the above diagram, the system should also include a non-infrastructure body which performs a number of important functions, as follows:

- Coordinating and facilitating the provision of expert guidance on the process and practice of data management. This service is provided to custodians and managers of research data, within and beyond those research infrastructure facilities funded by the Commonwealth.
- National coordination functions, including for accreditation processes (e.g., overseeing the organisations enabled to mint DOIs, curatorial standards, etc.)
- International coordination and linkages, e.g., the RDA.
- Support for development of national policy by government.

The necessity for such a body has been made clear by the demand for particular contributions of ANDS. Activities such as bringing together communities of data management practice, and providing insight and guidance into data curation for less-expert custodians, have been valuable and there is an ongoing demand.

A national data management body is therefore required—one which is empowered to co-ordinate and facilitate the provision of expertise in data management practice and process where it is needed, so enabling the good management and curation of data by the bodies with responsibility for it. Such a body would interface with parties including, but not limited to, those encompassed by the above diagram.