Submission Template

2016 National Research Infrastructure Roadmap Capability Issues Paper

Name	**Dr Greg Storr**
Title/role	**Chair of the Steering Committee of the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE)**
Organisation	**MASSIVE** **Coordinating Institution:** **Monash University**

Preamble

The Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) is a specialised High Performance Computing (HPC) facility that is a collaboration between four Partners: Monash University, CSIRO, Australian Synchrotron, and ANSTO, and two Affiliate Partners; the ARC Centre of Excellence in Integrative Brain Function, and the ARC Centre of Excellence in Advanced Molecular Imaging.

MASSIVE provides data processing and analysis services to drive research in disciplines such as biomedical science, materials research, engineering and geosciences – with particular focus on neuroscience and molecular imaging through partnership, and underpins a range of advanced imaging modalities, including synchrotron X-ray and infrared imaging, functional and structural Magnetic Resonance Imaging (MRI), X-ray Computer Tomography (CT), electron microscopy and optical microscopy.

With the goal of impacting the new generation of wet lab and experimental scientists who are capturing ever-increasing amounts of data, MASSIVE has intentionally taken an approach that is different from peak HPC centres. This approach is to underpin key data producing instruments, and make accessing HPC services easier for a broader cohort of researchers:

- MASSIVE has a dedicated instrument integration program that tackles peak data producing instruments and provides researchers with “in-experiment” data processing. Instruments supported by this program include the Imaging and Medical (IMBL) and X-ray Fluorescence Microscopy (XFM) beamlines at Australian Synchrotron, and the Ramaciotti Centre for Cryo Electron Microscopy Titan Krios Microscope.
- MASSIVE runs the Characterisation Virtual Laboratory (CVL), a NeCTAR-funded project that has evolved into a coordinated characterisation informatics program to make research computing more accessible to a new generation of scientists. MASSIVE undertakes this program of work in collaboration with project partners that include 6 research-intensive Universities, NCRIS funded capabilities and other organisations[1]. The environments, software and data management infrastructure developed under this program have scaled up significantly and the total capital value of the instruments being underpinned by this project exceeds $242M across 21 University, state and national facilities. Our usage metrics show that over 50% of researchers who are introduced to the CVL continue to use it.

By the definition provided in National Research Infrastructure Capability Issues Paper, MASSIVE is a Tier-1 facility: it has provided HPC computing and data processing services to over 1,000 researchers across over 100 institutions. However, MASSIVE is different from the other facilities that fall under this definition: MASSIVE is an **integrative HPC** facility, in that it has a specific focus on instruments and the new generation of wet lab and experimental scientists who require data processing capability.

Approximately 55% of the projects underpinned by MASSIVE are in the life sciences, with a particular focus on brain science and molecular science through partnerships. Over the last 3 years, requests for access to MASSIVE through the National Computational Merit Allocation Scheme (NCMAS) have been typically 3-4 times greater than what is available, indicating that demand is strong for the capabilities offered by MASSIVE.

MASSIVE is a critical part of the national HPC environment but currently lacks connection into the national funding landscape, even though it was seeded through funding from the NCRIS research infrastructure programmes (NCI Specialised Facilities - 2010). The facility has been in operation since 2010 and since initiation received $14M funding in total, of

[1] Monash University, University of Queensland, University of Sydney, ANU, University of Melbourne, and University of Western Australia, AMMRF, Australian Synchrotron, ANSTO, the National Imaging Facility, QCIF, Intersect, Research Data Services.

which $1.2M under the NCI Specialised Facilities program, $2.4M NCRIS project funding, $1.45M from the Government of Victoria and $9M from it's partners.

The partners and affiliated partners of MASSIVE, affirm:

- MASSIVE is a unique facility that is recognised internationally;
- MASSIVE is a critical part of the infrastructure strategy of its partners;
- Australia needs to invest in an **integrative HPC facility**, such as MASSIVE, that provides a nexus between instruments, new users communities, and data science techniques;
- There needs to be a more coherent approach to the national HPC landscape, and MASSIVE supports greater integration of HPC capabilities at all levels;
- Applied artificial intelligence is of increasing relevance to research and it provides researchers the ability to gain insight from data that may otherwise be unfathomable. A HPC facility that couples fast data processing with an applied artificial intelligence service is an important part of a future Australian HPC strategy; and
- Australia is now ideally placed to develop an informatics fabric that connects the hundreds of data producing imaging instruments (both NCRIS and University funded) with tools and researchers across Australian institutions.

The MASSIVE partners and affiliates are of the view that, subject to appropriate review and approval, NCRIS funding dedicated to MASSIVE would be rewarded with significant dividends for Australian researchers.

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

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

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

It is essential that Australia invest in developing easier and more powerful online infrastructure for science communities. These Virtual Laboratories (or "Digital Gateways") provide researchers with the ability to more easily leverage centralised IT infrastructure in a way that is specific and relevant to their community.

It is important to note that major European and US investment towards specialised online digital infrastructure will naturally outweigh Australian investment. For example, a sizable proportion of the €1B Human Brain Project (HBP) is devoted to developing sophisticated online platforms for neuroscientists to perform simulation experiments, and to access important data collections. It is natural that Australian researchers will increasingly want to use these internationally developed platforms through collaboration or other means.

Using overseas digital environments will undoubtedly provide Australian researchers with value. However, it is essential that Australians become the "Virtual Laboratory builders" as much as the "Virtual Laboratory users". There is significant benefit to the knowledge economy to being involved in the design, development and innovation that occurs when developing digital research platforms. For example, developing a cloud platform that allows psychologists to perform digital experiments on brain connectomics data is not dissimilar to developing a next-generation logistics routing platform for business – both require the integration of sophisticated modelling techniques, with big data, instruments, and user experience.

To participate in large-scale international endeavours, it is important that Australia provides the collaborative funding mechanisms required to participate in international programs, such as Horizon 2020 and its projects such as HBP. This collaborative funding mechanism does not have a natural home in one of the existing Underpinning Research Infrastructure projects, and therefore the opportunity to leverage funding to promote closer international collaboration is lost.

A single entity for Underpinning Research Infrastructure (NeCTAR, ANDS, RDS) would create a better opportunity to capture all of the informatics work being done for a given community (i.e. Characterisation) to establish international connection opportunities.

The timing is important: Australia is currently well placed to develop a strong national and international user base of online digital research platforms, based on national experience and expertise built through NeCTAR Virtual Laboratory program, and the Research Data Services A1.x program.

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

International consortium will develop significant initiatives that outweigh what any one country can undertake alone. In these circumstances, it is essential that Australian researchers are able to participate not just as users but also as collaborators and co-creators (as per our answer to Question 3).

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

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

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

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

Where funded by national funding, allocation of computing time on computing facilities should be through a national merit allocation scheme that is fully **independent** of the organisation(s) providing the computing platform.

An independent merit allocation committee is important for the purposes of good governance and transparency.

It is also important if Australia adopts a tiered model of HPC facilities, as done in Europe. In this case, the merit allocation scheme will become a primary method of directing projects to the most appropriate tier.

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

See question 10.

Question 10: What financing models should the Government consider to support investment in national research infrastructure?

It is important that funding models for digital platforms (such as Virtual Laboratories) are able to scale or adapt to accommodate their growth and scope. We propose that funding for virtual software infrastructure should be more dynamic, with the ability to increase funding for projects with an increasing user cohort, and the ability to decrease or shut-down projects that have not shown uptake and impact. To achieve this, it is important that projects are run and governed transparently and are required to transparently report on impact on research, including usage information.

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

HPC is increasingly an essential component of both industrial and clinical research workflows. It is therefore now becoming important to develop an Australian HPC service that adheres to appropriate international ISO quality accreditation standards.

Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?

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

Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

Health and Medical Sciences are increasing **data sciences**: scientists develop understanding by performing experiments across a wide selection of instruments and modalities. They then distil and piece together the clues gathered using data processing and analysis techniques.

The ability to process, analyse and visualise this data will be essential to Australia's future success in the health and medical sciences and should be a major driver of an Australian HPC strategy.

Approximately 55% of the projects underpinned by MASSIVE are in the life sciences, with a particular focus on brain science and molecular science through partnerships with the ARC Centre of Excellence in Integrative Brain Function and the ARC Centre of Excellence in Advanced Molecular Imaging, and partners Australian Synchrotron and CSIRO.

To deliver to the health and medical sciences research community, HPC must be delivered differently to existing nationally funded projects, and should shift to bring balance to the new emerging HPC demands which highlights:

- Usability over capacity;
- Hardware and software suited to data analysis over modelling and simulation;
- Underpinning large number of high performing wet and experimental laboratories, with growing data processing needs, over peak-scale simulation projects that are already well served by the major two national HPC facilities;
- Connectivity and workflows, including connecting with instruments such that data is processed, analysed and visualised automatically so that inexperienced users can begin to leverage new-generation instruments.
- HPC for life sciences must be porous and flexible: data must easily flow in and out, and flexibility in allocation of resources is important to allow for normal variations and unexpected results that occur during physical experiments and data collection; and
- The option of fully secure and ISO quality assured environments that provide strict data access control and security to support clinical and sensitive data.

Question 16: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

Environment and Natural Resource Management

Question 18: Are the identified emerging directions and research infrastructure capabilities for Environment and Natural Resource Management right? Are there any missing or additional needed?

Question 19: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 20: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Environment and Natural Resource Management capability area?

Advanced Physics, Chemistry, Mathematics and Materials

Question 21: Are the identified emerging directions and research infrastructure capabilities for Advanced Physics, Chemistry, Mathematics and Materials right? Are there any missing or additional needed?

Question 22: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 23: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Advanced Physics, Chemistry, Mathematics and Materials capability area?

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?

National Security

Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?

Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?

Underpinning Research Infrastructure

Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?

Users of Australia's characterisation facilities include a wide selection of researchers and industry, across medicine, the life sciences, engineering, materials science and geoscience. Since original NCRIS funding in 2004, these fields of research have evolved into **data sciences**, where data analysis sits alongside other core research methods such as experimental design and simulation. Scientists are increasingly collecting evidence, in the form of ever-larger data sets, across an ever-wider cohort of instrument and modalities.

Australia needs to underpin investments in instrument infrastructure with advanced computing:

To meet the data processing requirements of the next generation of Australian instruments and to underpin the large cohort of newly minted data scientists who will use them, it is important that Australia invest in an **integrative HPC capability**.

Over the past decade, there has been significant Federal Government, State and institutional investment in characterisation instruments. Whilst new technology has created immense opportunities for world-class discoveries, researchers are faced with an image glut[2] that is a significant hurdle to these discoveries. Australia's contribution to scientific innovation and knowledge will be influenced by our ability to help researchers analyse large multi-modal datasets. The current demand for computing capability that underpins

[2] http://www.nature.com/news/the-struggle-with-image-glut-1.19834

instruments is strong and will increase exponentially with the increased resolution and capability of detectors.

To meet the data processing requirements of the next generation of Australian instruments, Australia needs an "integrative" facility that is different to traditional HPC facilities:

The attributes of such a facility are:

- **A dedicated program to integrate instruments with HPC, thus providing scientists with the ability to perform fast and sophisticated analysis of captured data.**

  Data processing and analysis is increasingly a core real-time part of scientific experiments. An **integrative HPC** system supports large-scale analysis in real-time during an experiment which allows researchers to make real-time decisions about their experiment. In turn, this increases the return-on-investment (in both the experiment, and the underpinning infrastructure) by ensuring that resources are used in the most efficient way possible.

  A national **integrative HPC** system will underpin and de-risk investment in flagship data producing instruments (including existing and new instruments at ANSTO and Australian Synchrotron) and lift the capability of researchers using these instruments.

  By developing automated workflows, and capturing data from the point of generation, an **integrative HPC** capability will impact a very wide range of users. Significantly, it will provide data processing services to research groups that would not otherwise have processed data at the scale or with the robustness needed to best leverage the data generated.

- **Dedicated support and programs to on-board and support new researchers from fields of science that have not traditionally used HPC.**

  The uptake of advanced instrumentation by scientific communities is driving demand for computing services by fields of science that are new to the problems associated with big data. For example, the uptake of functional MRI techniques is creating significant demand for specialised MRI processing by the psychological sciences, a community that has not traditionally used HPC.

  HPC facilities increasingly need to support easier access – and this must extend beyond the expert command line interfaces – to build usage from amongst the wide cohort of experimental or wet-lab scientists. For example, explicit support for remote desktop and visualisation allows a range of scientists, including a large cohort of inexperienced HPC users, to access important tools and data. Across the CVL and MASSIVE, we have shown that over 50% of researchers who first try remote desktop access as their gateway to a HPC system continue to do so and is a powerful tool to build usage amongst the new generation of experimental and wet lab scientists.

- **A focus on applied data science, including applied artificial intelligence, to address the data challenges being faced by a broad spectrum of researchers – including wet and experimental research laboratories.**

  The confluence of big data, deep neural networks and tightly coupled parallel computing is enabling large commercial multinational companies and innovative start-ups to apply artificial intelligence across a wide range of problems, with increasing sophistication, insight and accuracy. Applied artificial intelligence is of increasing relevance to research as it provides researchers the ability to gain insight from data that might otherwise be unfathomable. **An HPC facility that couples fast data processing with an applied artificial intelligence service (including specialised hardware, software and expertise) will be a transformative part of a future Australian HPC strategy.**

> Partnering with industry will provide an accelerated way to achieve this, and the precedent to achieve this strong. Over the past 10 years, ultra-fast parallel data processing has been made possible by the hardware developed for the computer games entertainment industry. In the last two years, ever more sophisticated applied AI is being developed by large multinational Internet and semiconductor companies. Leveraging this industry investment will create significant rewards to researchers.

To on-board and underpin this new generation of experimental and wet lab scientists, a centralised facility is unlikely to achieve deep collaboration across key instruments and facilities. Rather, we recommend a hub and spoke model that provides peak capability and coordination centrally, and spokes of activity to integrate with instruments and specialised capabilities. This has been the central concept behind the coordination of the MASSIVE (hub) and CVL projects (spokes) and has led to strong engagement at both the peak data producing facilities (e.g. Australian Synchrotron), distributed capabilities (e.g. MRI facilities across Australia), and specialised capabilities (e.g. integrating Atom Probe techniques at University of Sydney with the Australian research cloud).

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

Australia needs coordination and leadership in characterisation informatics if it is to fully leverage investment in imaging infrastructure and computing facilities

Characterisation in the Australian context is a diverse community that is underpinned by a wide range of instruments. However this community is united by the need to address common informatics challenges. Whilst highly multi-modal and distributed, the Australian imaging community has successfully coordinated across key informatics initiatives (MASSIVE, NeCTAR Characterisation Virtual Laboratory, RDS A1.4 Image Publication, NIF Informatics). In many cases, these initiatives have been recognized as best practice and have underpinned hundreds or thousands of researchers. In all cases, expectations and key performance indicators have been met or exceeded. The effect of these projects has been transformative and hundreds of scientists rely on these projects and their outcomes on a day-to-day basis.

As a part of these initiatives, over 50 instruments have been integrated with cloud-based data management software to ensure that all data generated by these instruments is automatically captured, managed and delivered to the cloud for processing, analysis and visualisation in the Characterisation Virtual Laboratory (CVL), or MASSIVE, or data management tools hosted by RDS. The project is seeking to continue to double that effort and expand to new types of instruments and facilities in the next year. **The total capital value of the instruments being underpinned by the project coordinated under this endeavour is $242M across 21 University, state and national facilities[3].** Combined, this integration effort has underpinned over 1,500 researchers and many of those have gone on

[3] **Facilities participating include:** Animal MRI Facility, Florey Neuroscience Institutes, Australian Centre for Microscopy & Microanalysis (USydney), Australian Synchrotron, Biological Optical Microscope Platform (UMelbourne), Bragg Institute (ANSTO), Center for Advanced Imaging, (UQ), Centre for Microscopy and Microanalysis (UQ), Florey (Melbourne Brain Centre), FlowCore (Monash University), Melbourne Brain Centre Imaging Unit, MicroNano Research Facility (RMIT), Monash Biomedical Imaging (Monash University), Monash Biomedical Proteomics Facility, Monash Injury Research Institute, Monash Micro Imaging, Monash Micro Imaging (AMREP), Queensland Brain Institute, Royal Children's Hospital, St Vincents Hospital, The Clive and Vera Ramaciotti Centre for Structural Cryo-Electron Microscopy, X-ray Microscopy Facility for Imaging Geo materials (Monash University)

to be core users of MASSIVE and cloud-based virtual laboratories. Furthermore, the groundwork has led to the development of specialised workflows and toolboxes for next-generation modalities, including: CryoEM, atom probe, cytometry, and lattice light sheet.

Building on this, Australia is now ideally placed to develop an informatics fabric that connects **the hundreds of data producing imaging instruments, both NCRIS and University funded, in research institutions across Australia and their users**. The goal of this fabric is to allow each and every instrument facility to tell their users on the day of their visit:

"Your data is now being ingested by a safe nationally connected data management network and from there you can easily access it in an online collaborative environment that has many of the tools and services that you need to gain insight".

An initiative such as this would both provide informatics capability to the long tail of imaging users who are increasingly burdened by big data, underpin flagship instruments that have immense data processing requirements, and provide a platform for collaboration over both data and tools. At national scale, this undertaking requires national coordination and leadership to work across and underpin the networks of instruments, computing facilities, and user communities.

Developing a broad imaging informatics network across instruments has a number of very positive attributes:

- It is low risk. Investments in instruments have demonstrated both scientific merit and community need, and underpinning this investment with eResearch capability almost guarantees engagement;
- It increases return on investment and de-risks instrument investment; and
- It captures a large cohort of users, which has been demonstrated by the very large cohort of users underpinned by the CVL.

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

Other comments