Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Rob Cook**
Title/role	**CEO**
Organisation	**QCIF (Queensland Cyber Infrastructure Foundation) – see Attachment 1**

Note: This submission includes four attachments that detail QCIF and its six university members' major interests in issues covered by the roadmap capabilities issues paper. Specific items are highlighted in responses to the questions.

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

In addition to health and medical sciences, agriculture, the environment and other non-human health are large areas of nationally important research. QCIF works with digital infrastructure and bioinformatics for both health and other life sciences.

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

The characteristics are appropriate. The focus could be on needs, as well as benefits and outcomes.

QCIF favours governance structures that include a corporate-style Board responsible for strategy and direction, with advisory committees providing research and technology advice and stakeholder input.

QCIF is the organisation responsible for research informatics and digital infrastructure (eResearch) in Queensland. It has recently drafted a strategic plan for the coming five years with the strong support of all of its six member universities for what they regard as a highly successful model. This plan is based around close engagement with the national eResearch infrastructure plan and its implementation with its own Queensland-based governance. QCIF's plan and governance are described in Attachment 1.

For the governance of Data for Research and Discoverability see response to Question 35

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

Australia needs engagement with international data repositories and international research/science clouds, both to facilitate access and abaysis of international data

and for international research to access and analyse data in Australian repositories. Engagement will require membership in the governing international organisations and funding for participation. This is unlikely to be funded by any single institution and selected engagements should be funded by national infrastructure investments.

Health and life sciences, and the environment and natural resources both have rapidly developing global networks of data and computational resources and there is the intention to establish cloud links to Australia. ELIXIR in Europe and BD2K in North America are important targets for bioinformatics.

This is in addition to international communications networking and will require investment. The investment in ANDS has included Australia's seat at the table in international research data policy organisations. This should continue and should be extended to international research cloud developments.

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

No response

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

Research workforce skills are a serious impediment in the effective research application of data and computation infrastructure – see the section about Skills Development in Attachment 1. QCIF's members regard skills development as one of the most crucial issues in improving research excellence through the application of digital infrastructure (eResearch). This is a national problem and should be considered as a research infrastructure issue.

There is a severe shortage of skilled technical staff to build and operate the national research digital infrastructure, including research domain informatics (such as bioinformatics).

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

Investment in:

- usability of platforms, tools and methods for computation and data skills developed for research community use
- research community development of tools, methods and data collections could include investment in training materials
- carpentry-style skills development aimed at generic IT and tools skills
- skills development in data management and the effective application of (big) data
- cadetships for training graduates to power the technically demanding aspects of developing and operating new advanced approaches to the application of informatics and digital infrastructure to research fields

An approach to organising these investments would be the investment in a research sector peak body such as AeRO (the Australian eResearch Organisations) to coordinate the development and dissemination of the highest priority skills development and training materials.

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

No response

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

No response

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

No response

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

No response

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

Institutions and researchers already expect that research data repositories will meet production research benchmark criteria for trust including security, privacy, resilience and longevity of data, and QCIF is expected by its members to demonstrate that it meets those criteria. Industry, government and some research users (health for example) require these guarantees.

Repositories will be asked to demonstrate that they are meeting increasing levels of maturity for production research data facilities. QCIF expects that within five years they will required to undergo some level of certification.

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

No response

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?

No response

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

No response

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?

Bioinformatics is an increasingly important service for researchers in health and medical sciences and also in life sciences, agriculture and the environment. A national facility for bioinformatics such as those proposed by EMBL-ABR (University of Melbourne) and the Nectar Bio-cloud would focus existing separate regional efforts on the development of:

- national platforms (such as GVL, Omics, MyTardis and Omero) that can be used to process and analyse data by scientists and clinical researchers without a deep knowledge of information technology
- availability of software tools for data analysis
- data collections
- skills in the application of computation
- international engagement
- access to computation, data and data storage.

QCIF is a significant participant in both EMBL-ABR and in the Nectar Bio-cloud.

Bioinformatics and other areas that require national research investment, such as **multi-omics, personalised medicine, annotations and bio-imaging** are discussed in more detail in Attachment 2.

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

ELIXIR and EMBL-EBI in Europe and BD2K, operated by the NIH in North America.

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?

No response

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?

An important national infrastructure is a digital platform that enables the storage, linkage and integration of environmental and national resource management data across a wide range of disciplines associated with terrestrial ecosystems and with the studies of the populations and communities that inhabit urban and regional ecosystems. Such an infrastructure does not exist today and the many research communities in the field develop individual data and computation systems within their own disciplines. AURIN has a platform that is iikely to be an important contribution.

Guided by TERN, Nectar (through the Eco-Cloud development), the Nectar Biodiversity and Climate Change Virtual Laboratory and other groups associated with the tropics, with agriculture and with the environment, a coalition is being developed to stimulate the development of such a national infrastructure. This infrastructure will require national infrastructure investment.

There are many applications in areas of national and regional need including:

- the future of the Great Barrier Reef
- the development of Northern Australia
- disaster resilience
- social and community development in local government areas
- and many others are emerging.

These cross over into many of the areas discussed in the Understanding Cultures and Communities section.

The national and regional collaborators named above are working with QCIF (see **Attachment 1**) to seek funding for the extension of QRIScloud, its Nectar and RDS platform, to build linkage and integration between the domains, and to enable translation of research to application in many areas and regions and to enable collaboration between industry, government and research to address important system-of-systems issues. Please refer to **Attachment 3** for more detail.

The infrastructure necessary to form the platform includes, amongst others, methods and standards for data integration, storage, data access and methods for complex computation over data stored at different sites. This infrastructure is intended to be deployed nationally.

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

A number of opportunities have been identified in Asia and the South Pacific for conducting similar projects and delivering associated training in other countries – tropical research and its applications in South East Asia, and biodiversity in the South Pacific to name two.

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?

No response

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?

QCIF and its members are working with the National Imaging Facility, AMMRF, Monash University and others using Nectar and RDS funding for the CVL (Characterisation Virtual Laboratory) to develop systems to store, access and analyse images. The ImageTrove and Omero image databases are operated on QRIScloud and support the activities of a number of high profile institutions including the Centre for Advanced Imaging, the Queensland Brain Institute and the Institute for Molecular Biology. Use is being extended to other members and to research at hospitals and MRIs.

This collaboration has been very productive. Further development and operations investment from the national research infrastructure are needed to further improve the facility and for ongoing operations in Queensland and other locations.

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

No response

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?

No response

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?

QCIF is partnered with AURIN and Griffith University to extend the application of the AURIN platform in social, economic and population health applications in Logan and Gold Coast cities. There are prospects of much more extensive applications by combining these fields of work with that described under Question 18 and in **Attachment 3**.

This practical research application needs national research infrastructure investment to develop the AURIN platform, together with the University of Melbourne, and to research its further application in many areas of Queensland.

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

No response

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?

No response

National Security

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

No response

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

No response

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

No response

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?

The definitions of Tier 1 and Tier 2 computing in the issues paper are a little misleading since Tier 2 computers are often much less powerful than Tier 1 but still have the capacity to support multiple research communities – as an example Flashlite at UQ is a data-intensive Tier 2 HPC designed for multiple communities with very large data-intensive compute needs, and the 30,000 cores of the national research cloud might be regarded as a Tier 2 computer serving a wide spectrum of research communities.

The demand for Tier 2 and cloud computing is increasing rapidly for data intensive computing over the increasingly large volumes of data generated by research and the use of the dry-lab modelling and simulation approach to research. This demand cannot be met by peak computing facilities alone.

Tier 2 and cloud computing aimed at data-intensive computing needs to be located close to the data storage that hosts the large data sets that research computing

consumes and produces minimising the costs and time involved in transporting large data sets over contemporary networks.

Research communities are creating large data sets stored at multiple locations and used for a range of research applications. Rather than moving large amounts of data over networks with large, but still limited bandwidth, researchers are likely to conduct their research using computing distributed over the sites where the data is stored using the local HPC or cloud facilities.

These Tier 2 and cloud computing and distributed computation requirements are growing rapidly to meet research data processing needs and will need support from the national and regional research infrastructures.

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

No response

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

Over the timespan of the 2016 Roadmap it will become feasible and cost-effective to use commercial clouds for some research, and to lease selected capital computing and data storage infrastructure rather than purchasing it outright. Consideration should be given in the roadmap to the use of operational funding as well as capital expenditure for providing access to digital infrastructure for data-intensive research.

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?

Attachment 4 contains a more detailed response.

The rapid increases in needs for data storage for nationally important collections and for published data, for computing and data storage to meet research community needs and for research funded by merit-assessed research mean that the national infrastructure should continue to fund equipment replacement and growth for the centres currently funded through the Nectar and RDS Projects.

Contemporary research depends on computation and data analytics over rapidly increasing volumes of data, and the generation of data from modelling and simulation for use in further analysis and visualisation. This means that the most important aspects of (big) data in research are its discoverability, its access, ease of use for research, re-use, sharing, publication and dissemination. The ideal research data system should be designed and constructed to deliver data reliably and rapidly to research and collaborations, driven by research communities and their needs.

Details of data capture, management, metadata and storage are important as they contribute to the analysis and usability of data for research, and need to be designed as such in an ideal system.

Data preservation is a key element of a production research data storage system. Research communities and research administration requires data to be maintained for periods as long as 15 years, and nominated national and research community collections may be required for considerably longer. Consideration should be given to investment in production data preservation tools or services that provide assurances of the necessary longevity.

QCIF and its members would like to see the description of an ideal research data system and data life-cycle improved to better reflect the components and their importance to data storage and data use and analysis, and to provide more emphasis on the international interoperability of data and the systems used to deal with data.

UQ has developed its Medici campus storage cache system that is being adopted for use with QCIF storage services. Medici connects directly to instruments, campus computers and researcher storage systems. Data transferred to Medici is automatically copied at high bandwidths to QCIF data repositories ensuring that data is replicated safely and is available on-campus for use rapidly whenever required.

Maturing production data systems and making data available for computation need further national research infrastructure investment.

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

There are a number of international data capabilities including RDA and EUDAT, and data capabilities for specific research communities such as EMBL-EBI for bioinformatics. More are expected to emerge for other disciplines.

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?

In QCIF's view the existing ANDS, Nectar and RDS Projects have established and are funding the operations of effective but separate services including assistance for research communities to build platforms over the services. The disadvantage is that at least parts of the services function independently of each other.

In future research would be better served by replacing ANDS, RDS and Nectar by a successor Data for Research and Discoverability entity to drive the development of research data services in Australia, and particularly their engagement with and work for research communities. A governance structure similar to that suggested in the response to Question 2 would be suitable.

Other comments

If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20 page limit of submissions.

No response

Attachment 1 QCIF – Queensland Shared eResearch Infrastructure

1. Positioning Statement

QCIF's members have agreed to work together over the coming five years (2017-2022) to build on the QRIScloud digital infrastructure that has been established under the NCRIS program since 2011. Under this program Queensland intends to work with NCRIS successor programs, the Queensland Government and their own co-investments to:

- Focus on excellence in health and life sciences, and the environment and natural resources through informatics and digital infrastructure
- Translate research excellence into application in practical industry and government projects, particularly large system of systems projects
- Supply expertise in the use and analysis of large data systems to enable new research insights
- Expand QCIF's skills development capability
- Provide access to very large research digital infrastructure by consolidating member capacity at QCIF, building regional capacity, integrating with national research infrastructure and expanding its use of commercial services.

This member alignment to promote QCIF's capabilities is unique in Australia and will form a strong partner for the national eResearch infrastructure.

2. Queensland's Contribution to the National eResearch Infrastructure

Shared eResearch infrastructure funded by investments from the NCRIS program, the Queensland Government and the members. Operating a fee-for-service model for value-added services to research and innovation in the research, government and industry sectors.

Operates as a not-for-profit. Governed by a Board comprised of members and two independents. Members own all the assets funded through QCIF – QCIF can decide usage policy for the assets it has funded. Staff primarily seconded from the members, some contractors where skills not available from members.

Its four businesses cover the gamut of research infrastructure discussed in the Roadmap Issues paper:

1. Informatics expert assistance for priority research communities
2. Data expertise - hosting, management, discovery, access and analytics
3. Skills development – in informatics and the application of infrastructure
4. Access to digital infrastructure – regional (QRIScloud), national and commercial.

Serving the needs of research institutions, research communities that span institutions and the application of research into industry, government and the community at large.

For Queensland, and an integral part of the national digital infrastructure (through its Nectar and RDS activities) and connected internationally to serve its user communities.

QCIF's model and strategy are strongly supported by all of its member institutions and this support has been reaffirmed in 2016 through a joint strategic planning program and through the attached letters of support. Members are committing their financial support for the coming three calendar years providing the stability and longevity that is required for a research service provider.

All QCIF members regard the QCIF model as highly successful. It is unique in Australia in successfully coordinating state activities and delivering a comprehensive service to all its members while integrating with national infrastructure. All its members regard it as essential for their continuing research and research application success that QCIF continues as an integral part of the national eResearch infrastructure, through the Research Infrastructure Roadmap or otherwise, and that QCIF, representing the views of its members, participates in the direction of national eResearch strategy and its application.

QCIF partners with AURIN, as well as with Nectar, RDS, ANDS, NCI, AAF and AARNet from the NCRIS eResearch facilities, and with CSIRO and ODIQ (the Open Data Institute Queensland).

3. Informatics expert assistance for priority research communities

QCIF provides community specific informatics expertise in bioinformatics for health, medicine and the life sciences and for a range of communities in terrestrial, environmental, ecology, biodiversity, agricultural, regional, urban and spatial sciences.

Research communities need funding and support to continually improve the tools, methods and data collections that represent the digital infrastructure platform for community research excellence. Each research community (health and medicine, environmental etc) uses specialised tools and computational methods to extract results from data captured for their research and from collections of data available from prior research or from government and industry. QCIF partners nationally to develop and make available tools and data collections.

Community tools, methods and collections are developed globally and Australian researchers need access to, and to be able to improve on and contribute to international best practice. This means collaboration and interoperation with global research community development and Australian developments of best practice

tools in gaps of specific interest to Australia. QCIF partners nationally to access international collaborations.

Research communities contain pioneers who develop the next generations of research community digital tooling, and community researchers who may be less digitally literate and expect comprehensive but easy-to-use current best practice tools. Pioneers will often develop their own tools – community researchers need access to platforms and considerable assistance, consulting and training to be able to use digital tools. QCIF provides the necessary assistance, consulting and training.

Research communities can translate their work into application in industry and government, particularly when there are well-established research platforms that can be hardened and improved for application use. QCIF works with industry and government to establish opportunities for leveraging existing digital platforms to project research into application, forming a bridge for sharing and exchanging the underlying data collections, applying research methods and building effective collaborations.

3.1 Health, medicine and the life sciences

QCIF operates a team of bioinformaticians (currently eleven including those from its member universities) to construct and improve platforms, tools and data and to assist and train biologists and clinical researchers. These services are relatively mature.

The QCIF team presently partners to build bioinformatics platforms that are needed by the research community nationwide including:

- GVL – the Genomics Virtual Laboratory – a platform providing ease-of-use for complex genomics and other procedures (with EMBL-ABR, funded by Nectar)
- Omics – a data management platform for multi-omics data originally developed to assist the BPA Sepsis project (with EMBL-ABR and BPA, funded by RDS)
- CVL – the Characterisation Virtual Laboratory – a platform including ImageTrove and Omero for managing bio- and medical images (with Monash, funded by Nectar and RDS)

These platforms have large user groups and are or will be operated by QCIF amongst other sites to satisfy the needs of local and regional researchers. QCIF is working with Nectar under its recent Agility Fund grant and with EMBL-ABR to engage with the research community through a Bio-cloud, built on the Nectar research cloud and specialised to provide simple easy access to the spread of research community platforms, tools, methods and data. These new infrastructure developments require national funding.

The platforms are capable of and researchers are seeking considerable expansion and improvement. This activity will require funding, preferably on a national basis for a national research community.

QCIF provides consultancy, assistance and training services for community researchers, currently on a subsidised fee-for-service basis. This service needs to be expanded to cope with increasing demand.

QCIF provides translation and application services to extend its platform and consultancy capabilities to clinical research in hospitals and medical research institutions.

3.2 Terrestrial, environmental, ecology, biodiversity, agricultural, regional, urban and spatial sciences

QCIF and its members are planning a terrestrial informatics capability, similar to the successful bioinformatics group, to address the combined needs of a wide range of research communities whose needs overlap through spatial coordinates, climate change and other characteristics. QCIF is working with TERN, Nectar, RDS and AURIN as well as their members to build this capability. This advanced research infrastructure is described further in Paper 2.

These research communities have each developed their own specialised digital platforms, tools and collections but in a manner largely siloed from each other as pointed out in the Roadmap Issues paper. This silo effect reduces platform use for the increasing demands of cross-disciplinary research, and for applications of the research in industry and government. These communities often depend on external data collections, such as those from federal, state and local governments and disparate research platforms makes this more difficult.

QCIF plans to work with its partners including the Open Data Institute Queensland to develop platforms and tools that work well for each of these research communities. Work with Nectar under its recent Agility Fund grant will lay the groundwork for an Eco-cloud platform to provide researchers with easy access to a range of specialised tools. The platform and the specialized tools for individual communities will need national funding.

The platforms are capable of and researchers are seeking considerable expansion and improvement. This activity will require funding, preferably on a national basis for a national research community.

QCIF provides consultancy, assistance and training services for community researchers, currently on a subsidised fee-for-service basis. This service needs to be expanded to cope with increasing demand.

QCIF provides translation and application services to extend its platform and consultancy capabilities. There are large government and industry projects in Queensland that will benefit from the application platform and services envisaged.

4. Data expertise - hosting, management, discovery, access and analytics

QCIF is an expert resource for its members and for the national research infrastructure in capture, storage, management and preservation of data, and in its discovery, use, sharing and publication. QCIF intends to build its capability rapidly using its involvement with the priority research communities described above and to disseminate its data expertise much more broadly.

Data expertise has been acquired through QCIF's work with the national ANDS, Nectar, RDS and RDSI projects and through active ongoing work with its member eResearch Centres and their research communities that are pushing the boundaries of contemporary use and application of data, and particularly big data.

QCIF has collaborated in the funding of a tier 2 data-intensive HPC (Flashlite) located at the University of Queensland for the shared use of all QCIF members for the rapid processing of large volumes of data. The part of the Nectar research cloud at QCIF has large memory nodes specifically designed for processing biology, and later environmental, data with large in memory processing characteristics. In our view this will become required infrastructure for computing over data in many research communities, and one not well-served by conventional HPC and more suitable for tier 2 and cloud computation application.

QCIF provides consultancy, assistance and training services for community researchers, currently on a subsidised fee-for-service basis. This service needs to be expanded to cope with increasing demand.

QCIF provides translation and application services to extend its platform and consultancy capabilities. ODIQ and QCIF have partnered to provide consulting and training to industry and government to share expertise in using data and build capability with Queensland Government funding and support from industry.

There is much to be learned about the effective management of data as a research asset, and we can expect much of this learning to happen over the next ten years. Funding for infrastructure to acquire or build the required platforms, tools and services, and to interconnect data sources and compute over data will be essential to keep current in research. This advanced research infrastructure is described in more detail in Paper 3.

5. Skills development – in informatics and the application of infrastructure

Effective application of the digital research infrastructure will only be possible if the potential users have or can learn the necessary skills. This can be done both by improving the usability of the infrastructure so that the skill level require is reduced, and by developing the necessary researcher skills. Anecdotal and some hard evidence suggests that in general researcher skill levels are insufficient and that as a result many researchers do not gain as much advantage as they could from digital infrastructure. QCIF has found that even "carpentry" courses aimed at improving basic IT skills for research are always oversubscribed.

QCIF runs a scheduled program of skills development courses for researchers in:

- Basic IT and data skills – software and data carpentry
- Informatics skills for using research community tools – currently in bioinformatics
- Skills in using QCIF's QRIScloud digital infrastructure and tools.

Some of these are free, or subsidised by member institutions, and some are fee-for-service.

QCIF runs a program of similar material aimed at government research agencies and through ODIQ at government and industry. This program will become more urgent and important as members and external organisations engage more in translation and application.

QCIF is partnered with AeRO, the Australian eResearch Organisations, and with Nectar to contribute to and take advantage of online skills development material.

There is a lack of skilled professionals available to operate informatics and digital infrastructure services. People with these skills are in high demand both within the sector and in industry seeking to apply advanced digital infrastructure. QCIF is developing a cadet program to introduce fresh graduates and post-graduates into on-the-job training to grow their skills in QCIF and its members research businesses and then increase the population of skilled professionals available.

6. Access to digital infrastructure – regional (QRIScloud), national and commercial.

QCIF provides private resilient research compute cloud (capital funded by Nectar) and data storage and access (capital funded by RDSI/RDS) services as part of the national research digital infrastructure, and extensions of these services that have been funded by QCIF members. The. Queensland Government and QCIF members have co-invested to develop and fund part of the operations (data centre and staffing) of the services.

QCIF share in the funding of tier 2 HPC clusters at its members, including the Flashlite data-intensive HPC at the University of Queensland. These QCIF shares are accessible by members. QCIF also offers a cluster-as-a-service on QRIScloud and access to a QCIF share of the NCI peak computing service.

QCIF cloud services are branded QRIScloud, and QRIScloud compute and data services are integrated by design from a user point-of-view, and integrated with member HPC services.

Operating costs of QRIScloud are shared between members and the Nectar and RDS national projects.

All QCIF QRIScloud and member/QCIF HPC services are heavily used. QRIScloud and some of the HPC are located in highly-rated commercial production facilities.

Every QCIF member has signaled their intention, as part of the current QCIF strategic planning process, to expand their use of QRIScloud, to reduce their own research digital infrastructure over five years instead preferring to fund QRIScloud to provide services on a more cost-efficient, resiient and trusted basis. This is on the basis that QRIScloud continues to be fully interconnected into the national research digital infrastructure at physical, virtual and governance levels and retains a position of influence.

QCIF and its members expect that NCRIS or its replacement will continue to provide direct digital infrastructure funding, both capital and operating, for shared national facilities to meet the needs of research communities and shared data collections, and that this funding will anticipate rapidly growing demand for data and for compute over data. QRIScloud is designed as a component of these ongoing nationally funded facilities.

QCIF recognises that commercial cloud facilities can meet some research needs now and will be able to meet increasing amounts in future years. QCIF aims to provide an integrated service that offers researchers commercial, private and hybrid cloud services with a differentiated cost structure. Bursting into commercial services when QRIScloud resources are temporarily full will allow QRIScloud to offer an apparently elastic data and compute resource.

QCIF aims to offer a private cloud platform that supports the translation and application of research into industry and government R&D projects. This is aimed at data interchange between research, industry and government partners, and to make industry and government data more readily available to research. A second benefit will be collaboration between research, industry and government in solving complex system-of-systems projects from the real world. Paper 2 describes some examples.

Attachment 2: Infrastructure for Health and Life Sciences Research

QCIF is delivering bioinformatics capabilities to Health and Medical Sciences and to other areas of life science including agriculture and the environment. QCIF, on behalf of its members, is partnering with EMBL-ABR at the University of Melbourne to join a national bioinformatics consortium to foster international bioinformatics collaborations and to build national bioinformatics platforms, tools, data collections and skills development as well as providing access to digital infrastructure for researchers to use the facilities that are developed.

Bioinformatics is a critical resource for research in these fields and needs to be funded through the national research infrastructure investments, as well as state and institution co-investment.

QCIF is partnered with EMBL-ABR in the continuing development of the GVL (Genomics Virtual Laboratory) platform of workflows, processes and tools that is being continually expanded using NCRIS (Nectar, RDS, BPA) grants to support a wide spectrum of researchers. QCIF operates a GVL server on QRIScloud. GVL is already used internationally. It is important that support for GVL development and operations continues.

QCIF has partnered with the University of Melbourne, and Intersect in NSW, using NCRIS RDS funding to develop the Omics data platform that gathers, integrates and provides multi-omic data in support of the BPA Sepsis pathogen project. This ground-breaking work is designed to be expanded to support other multi-omic developments and it is important that it continues to receive investment from the national research infrastructure.

This joint work is engaging in a collaboration with Nectar using an Agility Fund award to engage with the research community through a specialised bio-cloud built on the research cloud and designed to provide the research community with the access that they require to platforms, tools, methods and data, all with a researcher-friendly interface.

The partnership between QCIF and EMBL-ABR has been very productive in building platforms, providing access to tools from elsewhere and to key data collections, and in operating platforms for the benefit of an extensive cohort of researchers.

These new infrastructure developments require national funding.

Multi-omics: In the coming years, more and more complex biological research questions will be addressed through multi-omics approaches. Assessing one single type of omic data (e.g. genomic or proteomic) provides an incomplete understanding of the biological system under investigation and currently limits researchers. However multi-omics based research is in its infancy and researchers are facing a number of new challenges:

1. The large volume of data, intrinsic to some omics but also linked to the number of samples and replicates involved in such studies.
2. The variety of data, with various features of each omics stream requiring different type of information encoding, file formats, and tools to process and analyse.
3. The difficulty of performing an integrative analysis of highly complex biological systems with a lack of existing methods and tools.

The research infrastructure will play a critical role in alleviating those 3 key challenges:

- More storage: high capacity and fast access for storage of omics datasets (raw and processed.
- More compute with large memory nodes that can support complex integrative analysis
- Platforms to manage and access data

Personalised Healthcare: The ability to identify the relationships between genomics information and phenotype will change the way medicine is performed and treatments will be tailored (choice of drug and dosage). Over the next 5 years, there will be a high demand for:

- More storage: high capacity, high security, high for storage of whole genome information and phenotype data, including health records
- More compute: population based studies on hundred/thousands of individuals
- Platforms to manage and access data

Annotations: The study of non-model organisms is critical for environmental and agricultural science. However a few tools are available to annotate such organisms

- More integrative analysis and systems biology to identify function.
- More compute for homology based methods
- Platforms to access a wide range of tools and data required in order to annotate gene function.

Imaging

- More storage: high capacity, high security, fast access for storage of images, including MRIs
- More compute, with an increased demand for GPUs for processing and deep learning.
- Platforms to manage and access data

Attachment 3: Infrastructure for Terrestrial Research and its Application

Environment and Natural Resources Management (Roadmapping Issues Section 6) includes a number of research communities with interests linked through the land, climate and other keys – terrestrial, environmental, ecology, biodiversity, agricultural, regional, urban and spatial sciences amongst others.

QCIF's partners are interested in many practical applications of these community's research, such as the Great Barrier Reef water quality, the development of Northern Australia, rework on the SE Queensland Plan and disaster resilience. Each of these require a system of systems approach to collecting data from a wide variety of sources and to creating models of individual systems that interconnect to form complex joined-up models. The outcomes will inform Queensland's agriculture, mining and environmental development.

QCIF is well-positioned to expand its QRIScloud platform to provide research production support for cross-disciplinary data exchange and collaboration given the appropriate investment. The result would be a platform for supporting the translation of multi-disciplinary research working with industry and government to tackle the practical applications described above.

Although the applications listed address Queensland-based issues the approaches can be developed and applied nationally. State and local governments and industry are considered likely to be interested in co-investing (in Queensland the Advance Queensland program is available).

System of systems modelling and data analytics is likely to require access to many individual data collections captured for distinct research or commercial purposes by different research domains or companies. This data is almost always developed in silo's for specific purposes and needs interconnections to create a data platform that can be used for accessing data from many sources for a single project.

A platform that provides the procedures and standards for creating and managing data from different terrestrially related sources would form an essential underpinning piece of national infrastructure for such purposes. The platform would need to be connected and interoperated with international platforms as they are developed.

Practical applications such as those mentioned above will mean extending production digital infrastructure developed for research into a platform that can support the application of research by industry and government. This will involve collaboration and the exchange of data between all the organisations involved in large system of systems projects, and the ability to build and use models and data analytics amongst them.

QCIF, TERN, the Biodiversity and Climate Change Virtual Laboratory from Nectar, spatial informatics from QUT and the tropical data hub from JCU and their partners

are planning the necessary components of underpinning digital infrastructure. The platform built by the AURIN project goes some way towards meeting these requirements and might form part of an ongoing development.

It will involve the hardening of current platforms to meet applications requirements, the extension of current research networking to applications projects and the development of new platform components to meet the needs of research/commercial projects.

Attachment 4: Infrastructure for Data and its Application

Data for Research and Discoverability

QCIF is building a research data infrastructure, with RDS, Nectar and ANDS, that provides access to data for active research usage, data for sharing for collaboration, and data for reference, publication and re-use. Research communities and researchers are able to guide the way that this is done for their preferred patterns of usage. Data can be accessed rapidly by HPC applications, by processes running on cloud virtual machines or shared through cloud-based portals and by users from their desktops.

The data storage system is flexible in that data can be accessed in different ways depending on the usage, and it can be discovered through national, institution, community or QCIF indexes.

Researchers need advice and training on working with data, particularly big data,, but more importantly on techniques for data analytics and statistics and for computing effectively with data.

A research production data system needs to protect data from loss or unauthorized use, to guarantee its preservation for its lifetime and to offer a trusted service that meets all the needs of research institutions including human clinical research.

Future data infrastructure investments need to recognise this complex environment and provide for data to be operated successfully within it. Data is an asset – it needs to be treated as such.

The Ideal Research Data System

There are two aspects to research data – data storage and management, and data access and use. It would better for understanding of the system if these were separated.

In an ideal research data system data custodians and data users would be able to use identical simple systems to store and access data whatever repository is used to store the data. Custodians would be able to assess the level of risk and the benefits in using any particular repository and choose one that meets their and their institutions' needs.

There needs to be considerable investment in the data infrastructure to bring its maturity up to this level.

Data Storage and Management

At each step metadata is expanded to reflect the provenance of the data. Rich metadata reflects the trusted status of the data itself. The data repository has its

own trust status based in its resilience, durability, longevity, security etc, and their trustworthiness can be credentialed.

- Data planning
- Data registration, space allocation and metadata creation
- Data capture, movement and ingest – maybe output from a computation process
- Data management and curation
- Listing in data catalogues
- Data access control – data may have an open, shared or private data licence
- Data archival and preservation
- Data discard – metadata retained

These activities are the responsibility of the data owner or data custodian as these are defined in the metadata. They are usually conducted using a series of tools associate with the data owner or the data repository.

Institutions may have policies about the location of data. Owners contract with data repositories to host data and to manage it and its access.

These steps with appropriate data management steps assure that the data can comply with the FAIR principles (findable, accessible, interoperable and re-usable)

Data Access and Use

Data that cannot be used an accessed and used is not worth storing. Usage methods and steps are:

- Data discoverability – though a catalogue, a research community, institution or repository portal or otherwise
- Data access – manually or programmatically - through user authentication and authorisation
- Data verifiability – through its trust status and provenance
- Data movement and copying
- Computation over data and data analytics
- Data sharing and collaboration – through a portal or access by a group of collaborators
- Publication and subsequent citation
- Conversion to a reference collection
- Data federation and integration
- Use and re-use in applications other than the original project

These activities are driven by data users, research communities, virtual labporatories or librarians who may or may not be related to the owner/custodian. They happen as a result of a data using tool through the cloud or as a result of other computation activity.

Computation over Data and Data Analytics

Research communities are developing standard workflows, processes, tools and methods that allow researchers to run complex processes over data. These processes are based on clouds to run the workflows and some of the tools and methods, and HPC to handle some of the compute or data-intensive sub-processes.

Such processes (compute and analytic) involving large amounts of data work best in an environment where cloud, HPC and data are co-located with very fast networks to carry data between the sub-systems. For processes involving large amounts of data stored in several locations in may be best to organise the computing so that it can be undertaken in parts using computers close to where the data is stored. This may be problematic for data stored remotely from research compute – such as data from government agencies and from remote reference collections including those stored overseas.

Future infrastructure investments can seek solutions.