Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Andrew Lonie**
Title/role	**Director**
Organisation	**EMBL Australia Bioinformatics Resource**

Submission Details
Author: Andrew Lonie, Director & Vicky Schneider, Deputy Director
On Behalf Of: EMBL Australia Bioinformatics Resource (EMBL-ABR)
Organisation: Hosted at the University of Melbourne
Type of Organisation: Not for Profit
Address: Lab-14, 700 Swanston Street, Carlton, VIC, 3053
Email: alonie@unimelb.edu.au
Phone: 0406 487 902
Website: www.embl-abr.org.au

Declaration of Interests
Andrew Lonie is also the Director, Victorian Life Sciences Computation Initiative (VLSCI) which is funded by the Victorian Government and contributing institutions and hosted by the University of Melbourne. This petascale facility delivers expertise and systems for life science computing.
The EMBL Australia Bioinformatics Resource ABR) is hosted at VLSCI through a funding agreement between the University of Melbourne and Bioplatforms Australia and exists as part of EMBL Australia's Associate Membership of EMBL in Europe.

Response Preparation
Contributions to the submission below were canvassed through the Australian Hub of EMBL-ABR across the existing network of node 'partners' around Australia. The response was led by Assoc Prof Vicky Schneider. Full list of node partners can be found at the end of this document.

This response is on behalf of the EMBL Australia Bioinformatics Resource (EMBL-ABR). The focus is primarily on life science infrastructure.

This era of big life science data has led to a consensus view that, even in Europe and the USA, institutions and countries cannot contain and manage all their data needs internally. There is a growing imperative to build strategic partnerships and extend collaborations to receive global exposure for their researchers' best data and assets in exchange for access to the best global data and assets. Both industry and academia now expect the infrastructure to support such exchanges to be facilitated efficiently. In Europe, the EBI has spawned ELIXIR, a federated model of engagement now extending beyond Europe, coalescing efforts across training, compute resources, international standards, tool development, platforms and data aimed at avoiding duplication and waste. In the USA, the Big Data to Knowledge (BD2K) program is addressing the same needs and working with ELIXIR to see that these efforts too may coalesce. Other efforts are building in other continents.

These are well-funded programs which Australia cannot reproduce at even a fraction of the scope. We have no choice but to align with these international efforts to ensure we have the skills and capacity to curate, package and add value to the Bioinformatics resources we have, making them findable and accessible through these larger initiatives and, in turn, return the value to the academy and industry in the form of easily accessible data and tools, and eligibility to participate in major, funded, collaborative research efforts.

In terms of management and mobility of the bioinformatics resources (from data, to tools, compute, platforms, standards and training) Australia's needs have only increased given the vast amounts of data across a variety of domains (e.g. genomics, proteomics, imaging etc) generated by publicly funded biological and medical research activity. Having a coordinated process to integrate and maintain bioinformatics resources across Australia and enabling its users in academia and industry is crucial to the life sciences and medical research.

What might this model look like? How would it work? As for many of the Bioinformatics initiatives happening at national level in several counties (BD2K, CyVerse, deNBI) as well as pan-national level (e.g. ELIXIR), an Australian Bioinformatics Hub channels activities to and from Nodes to join existing efforts across all of our institutions to ensure they are fully supported to engage with the world's data. Such an Australian Bioinformatics Hub would coordinate and provide national leadership through a consultative governance structure. EMBL-ABR is already implementing aspects of a federated model of engagement to enhance bioinformatics capability and responds to this issues paper from that perspective and experience to date.

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

The nominated capability areas contain within them the well-documented need for increased bioinformatics capability across the life sciences. Here we define Bioinformatics as the discipline that encompasses all aspects of the data life cycle. Bioinformatics is the science of storing, retrieving and analysing large amounts of biological information. It is a highly interdisciplinary field involving many different types of specialists, including biologists, molecular life scientists, computer scientists and mathematicians. Bioinformatics also comprises computational structural biology, chemical biology and systems biology (both data integration and the modelling of systems).

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

A federated model of resourcing for bioinformatics capability as currently being pursued by EMBL-ABR and reflecting successful models of national bioinformatics adopted in Germany (deNBI), Spain (INB), Sweden (BILS) to mention a few, and pan-national Bioinformatics (Europe's ELIXIR), requires the hosts and partners (e.g. matching funds between government funds and participating institutions (aka EMBL-ABR nodes)) to make significant cash and in-kind contributions to the development, maintenance, and operations. This creates an appropriate role for the Federal Government by offering stimulation for cultural change. However, this does create issues related to funding models, expectations, consistency of delivery of National services across a distributed infrastructure community, expectations, potential "national only or merit" use policies, confusion related to representation to the research community and within the university partner. An additional Governance principle could explicitly recognise the partner relationships and responsibilities and the additional value / enhancement they provide to the base capability and clarify the scope of responsibility. The recognition at the Federal Government level is crucial to the success of Australia's research outcomes and potential and, per se, a driving force that stimulates change and innovation in Bioinformatics to lead at a national and international level. Resourcing of a federated bioinformatics capability also entails provisioning of consistency of delivery of National services across the distributed community, through a clear set of the roles and responsibilities define with the governance principles of such initiative.

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

Yes, it is of vital importance to ensure that national capability areas have the potential to build on and engage with international solutions already developed and proved. This is particularly important in the area of ***managing, exploiting and safeguarding the increasing volume of data being generated by publicly funded research****. Many of Australia's agriculture, life sciences, medical and health researchers already participate in international programs, and it is increasingly more important for our researchers to be linked to global initiatives that are driving best practice and shaping the future of data-driven research in health. Australia holds a clearly recognised position in genomics research, and therefore we must ensure that a responsibility of national infrastructure is to build links and opportunities with international facilities, to support our researchers and to demonstrate our leadership role in the community. An example of international activities that are extremely relevant to Australia in this area, and key for* ***Australia's maximisation of impact in bioinformatics and actual life sciences, agri-food and medical research*** *outputs are ELIXIR https://www.elixir-europe.org/ and Big Data To Knowledge*

*(BD2K) https://datascience.nih.gov/bd2k. The ability to being able to sit as an equal with these programs is fundamental for Australia, making it not only an **early adopter** of the technologies, solutions and application in the data driven era, but also a partner in the shaping of the future in this area of data infrastructure.*

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

*In the area of bioinformatics and life sciences data infrastructure, there are a number of significant, **multi-year international programs** in place such as **Cyverse (NSF) and BD2K (NIH) in USA and ELIXIR in Europe**.*

It is critical that Australia capitalises on the advancements and solutions developed by these international efforts and, as appropriate, migrates the services and resources that exist internationally into Australia.

*This requires appropriate investment in both access and the development of national facilities that act as a gateway internationally and a national portal: providing uplift for the national infrastructure as well as providing a mechanism to ensure coordination at a national scale that is aligned with international programs and best practice. Aligning with programs such as ELIXIR will ensure that there is a **bidirectional pathway for the exchange of knowledge, reduce duplication and promote the adoption of Australian tools and techniques**. Some direct examples of how this will benefit researchers are illustrated in the following scenarios:*

Scenario A: when it is impractical to mirror critical data resources and associated services locally, for legal or logistical reasons. For instance, The Cancer Genome Atlas is 1PB of data with associated tools and services; 100,000 genomes UK.

Scenario B: where there is no clear advantage to implementing an Australian resource over access to international resource, or when equivalent benefits can be gained from an investment in international facilities at reduced cost. Again, TCGA; CyVerse; ELIXIR.

Scenario C: When Australia can contribute and benefit as a partner in larger international infrastructure consortia, gaining access to global resources for a local investment: eg ELIXIR.

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

*YES, this is **imperative in bioinformatics and absolutely necessary for Australia** to keep up and contribute to the developments, implementation and adoption of cutting edge, effective bioinformatics solutions. The life sciences research workforce is both consumer and producer of data infrastructure. The optimal impact and return on investment of the research infrastructure can only be achieved if researchers have the required skills to use and contribute to the infrastructure. Up-skilling researchers in data literacy and in tool development is thus a vital requirement that the research infrastructure must address. It is crucial to invest and support the fostering of competencies in data management, data analysis, resources (database, metadata collection platforms etc, annotation, curation) and tools development, including user experience (UX) and front-end design for bioinformatics platforms for life sciences research and its translation to **medicine, agriculture and bioindustries. Usability of bioinformatics solutions** is key to the realisation of the investments in this area and much needed for the impact of life sciences research.*

Hardware and facilities are one part of research infrastructure. Without appropriate expertise and skills, it is not possible to operate infrastructure and the benefit to researchers is significantly diminished by insufficient skills to analyse the data produced by such hard infrastructure. This is a defining characteristic of life sciences related research infrastructure, which is increasingly digital, generating enormous numbers of raw observations and pushing the research infrastructure bottleneck from data access to expertise in data analytics and informatics.

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

Training and skills development in bioinformatics across life sciences, agri-food and medical research is something that would benefit enormously by a nationally coordinated approach that maximises knowledge exchange, shareability, scalability and long term sustainability of postgraduate and beyond training in the core competencies which are key to the success and advancements of biological and medical research. Without a doubt, translating big data to knowledge presents several training challenges. Adherence to data-sharing and management guidelines and preparation of FAIR (Findable, Accessible, Interoperable, Reusable)-compliant data-sets being one to start with, as well as discovery and use of appropriate data repositories and storage systems. Another key area is seamless exploitation of cloud and virtual resources for effective training. Ensuring scalability and quality standards are critical concerns.

An increasing number of Australian universities and other research organisations are leveraging cloud infrastructure to deliver bioinformatics training to undergraduates, postgraduates and industry. A national approach will ensure that these training materials can be more easily shared between institutions, with adaptations to local needs being built on top of a standardised and shared core.

Resourcing the development of bioinformatics expertise which is focussed on international standards and world's best-practice will ensure the most efficient and effective outcomes for the communities who are to benefit from this investment.

If one accepts the premise that skills and training are a part of national infrastructure, then building coordinated national training programs around and within national infrastructure investments is an obvious way forward. As an example of how training and infrastructure can work together, the Victorian Life Sciences Computation Initiative (VLSCI) provides critical support for Victorian researchers through a combination of data analysis, compute resources, and postgraduate student training. The role of the VLSCI as the defacto hosting department of the University of Melbourne's MSc (Bioinformatics) program, through the synergy of the collaborative activities of the VLSCI throughout the precinct and the placement of the MSc (Bioinformatics) students, has been a critical factor in driving the success of that program. Leveraging researcher connections throughout the central precinct around Parkville has enabled VLSCI to create the excellent training and education environment necessary for building and sustaining the most successful bioinformatics program in Australia. This model of infrastructure facilities engaging in University-based education and training programs is a practical one in many instances.

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

Institutions have a role in up-skilling researchers in the context of a coordinated national and international approach. It is crucial researchers and technical specialists are upskilled to design,

maintain, develop, support and leverage the national infrastructure. However, they can only do this in the context of supporting a ***coordinated national and international approach****, which is why it is critical, at least in the life sciences,* ***that the national infrastructure provider take a leadership role in this activity.*** *Research institutions play a key role in identifying infrastructure training needs and can support roles that engage with the development of training solutions that meet their requirements****.*** *Similarly, these should be responsible to ensure that Postgraduate students and ECRs have appropriate skills by encouraging attendance to training events and providing training scholarships. Crucially, making data literacy skills a compulsory requirement of the PhD program.*

Research Institutes therefore also play a significant role in the dissemination of best practices across the data life cycle, which actually rely in having appropriate data infrastructure. Imagine a researcher wishing to access a particular dataset who might also find tools which have been used by others to help them interrogate that same dataset – thereby cutting out some of the duplication going on across research projects currently. Although it sounds trivial, significant effort still needs to be made to facilitate such process.

The scale of the problem was outlined by BD2K's Dr Vivien Bonazzi in an EMBL-ABR interview where she explained that their Data Commons project is tasked with approaching multiple NIH Institutes and centres to seek out the most highly-used digital artefacts (ie. all digital artefacts generated during NIH-funded research, including data, tools, workflows, documents, etc). From this 'stocktake' they house these artefacts in repositories and libraries where they are more accessible and available to researchers: on the cloud, through APIs or through publicly available indexes.

This was also endorsed recently in an @Big Data Expo interview with Dr. Liz Worthey, Director of Software Development and Informatics at the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama where she sees research institutions as an essential participant especially in the clinical application of data, not just for training, but as a vital intellectual resource from which to get the value out of that data: " we have … a lot of algorithms that you might use only exists in the brains of MDs, other clinical folks, or researchers. There is really a lot of human computer interaction work to be done, so that we can extract that knowledge", she said.

Researchers trained at the outset in an understanding of the full data life-cycle are essential to maximising value from publicly-funded research data. Sharing of data increasingly means sharing of the intellectual assets used to derive meaning from that data.

The ***national infrastructure provider ensures that local infrastructure is aligned with and supports the longer term and national objectives.*** *The importance of having a direct participation in the development of infrastructures lies in the key role of ensuring such infrastructure matches the institutional needs and effort is used at maximum efficiency, maximising sharing, interoperability and portability of bioinformatics solutions, and is a measure of the* ***co-investmen****t in the national infrastructure. The infrastructure provider ensures that the developments can be part of a larger national level networks, where knowledge exchange, expertise and actual pragmatic resources and tools are shared.*

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

Our only comment here is that whatever the principles developed in terms of the process adopted (merit based vs partial cost recovery vs total cost recovery), they should ensure that the data and

*tools adhere to similar efforts **internationally,** so best practice is adopted throughout. As an example adhering to FAIR data principles (doi:10.1038/sdata.2016.18), interoperability and adoptability of standards. While we accept that the underlying infrastructure is not limitless, we would promote as broad an availability to the research community as possible**.***

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?

It is expected that any activity operating as an EMBL-ABR activity would adhere to international standards as expected by the life sciences community, particularly with regard to training.

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

In the area of life sciences data infrastructure and application platforms there are a number of key exemplars that have been developed at national levels. In the USA, the national approach (BD2K) is being led through the National Institutes of Health and CyVerse and there are other programs being developed across Asia. In Europe, these national efforts are linked and catalysed by ELIXIR https://www.elixir-europe.org/, the Spanish Network of Bioinformatics Infrastructure (INB), The Swedish Bioinformatics Infrastructure for Life Sciences (BILS), the Dutch Techcentre for Life Sciences (DTL) among many others. Here we describe in particular the German Network of Bioinformatics Infrastructure (de.NBI). de.NBI was established to coordinate bioinformatics service provision across Germany. It brings together expertise and resources through dedicated service nodes in areas such as microbes, integrative bioinformatics, data management, crop bioinformatics, human genomics and proteomics. de.NBI is coordinated from a hub hosted at the Bielefeld University and includes over twenty partner institutes across Germany. It is national infrastructure supported by the Federal Ministry of Education and Research providing comprehensive, high-quality bioinformatics services to users in life sciences research and biomedicine. Besides the service provision in areas such as tools, platforms, data, compute, de>NBI also plays a role in coordinating bioinformatics training provision at national level. The nodes organize training events, courses and summer schools on tools, standards and compute services provided by de.NBI to assist researchers to more effectively exploit their data.

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 at the heart of modern biology *and with the developments in high throughput technologies that brought* ***clinical and personalised genomics*** *to a practical reality, it needs to be properly supported and resourced. Large-scale sequencing is a core part of today's clinical studies, to for example, facilitate accurate diagnosis of and discovery of new genetic variants underpinning disease.*

Bioinformatics is a fundamental platform that spans the range of emerging directions identified in the Issues Paper. In the area of ***big data linking data from genomics i****s critical to the accelerated advancement of trials, clinical practice improvement and health innovations. The world of 'omics' is driving new approaches to human and animal health and agriculture. Ensuring that the research community have a****ccess to the data infrastructure and tools to explore, exploit, share, collaborate, and develop solutions in a reliable, robust, repeatable environment*** *is fundamental to ensuring and improved outcomes for the community, maximising the benefits obtained from individual research programs for the collective benefit, and leveraging international programs and opportunities.*

It should be noted that with the de-commissioning of the Blue Gene-Q at the VLSCI by the end of 2016, Australia will lose a significant life-sciences focused supercomputing capability and a resultant reduction in research HPC capacity. Alternative HPC capability is being investigated, but now would not appear to be the time to disinvest in this level of research infrastructure.

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 - a distributed infrastructure for life sciences information *- ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states and enables users in academia and industry to access vital data, tools, standards, compute and training services for their research. https://www.elixir-europe.org/ . Through its EXCELERATE program, it currently funds ~19M Euro of activity with 41 partners in 17 countries.*

EMBL - EBI (European Bioinformatics Institute) *- based in the UK, it is Europe's premier institute for the support of bioinformatics and life sciences research, technology development and transfer, training and services to the scientific community. EMBL has 21 member states and two associate members (including Australia). It has a staff of ~520 and an annual expenditure of around 60M Euro.*

Genome Canada *will invest around CD$165M in 2015/16. It is a central Canadian funding agency that invests across all sectors of the Canadian economy, including health, agriculture, aquaculture and mining. Over half their investments have been in health in collaborative projects with universities, industry and research agencies and has invested over $1B over the past 10 years.*

Genomics England *– 100000 Genome Project, will invest GBP300M over three years. This is a consortia between the NHS, Health Education England, Public Health England and 73 NHS Trusts and Hospitals to sequence and analyse 100,000 genomes to transform the way health-care is delivered. The Garvan Institute participates in this program.*

The National Institutes of Health (USA) Big Data to Knowledge Program (BD2K) *invested around USD$32M in 2014 with a total investment of USD$656M by 2020. The Program is designed to enable biomedical research as a digital research enterprise to facilitate discovery and support new knowledge and to maximise community engagement. Thirteen Centres of Excellence across the USA.*

A fresh initiative that has just started and is again embedded in ELIXIR (who acts as the coordinator) is CORBEL, CORBEL - Coordinated Research Infrastructures Building Enduring Life-science Services https://www.elixir-europe.org/about/eu-projects/corbel, which aims to create a platform for harmonised user access to biological and medical technologies, biological samples and data services required by cutting-edge biomedical research. CORBEL will boost the efficiency, productivity and impact of European biomedical research. CORBEL actually follows on from BioMedBridges - the first cluster project bringing the biomedical sciences research infrastructures together - and will build on its achievements.

While not all of these investments are completely focussed on the creation and operation of infrastructure, they provide an insight into the scale and focus being undertaken in the area of life sciences data infrastructure and the importance that has been placed on this area across a number of jurisdictions.

Since 2015 Australia, through the EMBL Australia Bioinformatics Resource, has been actively engaged in the development of potential opportunities to collaborate with ELIXIR and as a bilateral opportunity.

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?

In the category of National health and medical big data capability, covering:
o infectious disease outbreaks
o bioinformatics skills,
the definition of bioinformatics skills needs to be given the broadest possible definition as the needs are core across undergraduate, post-graduate and senior researcher levels in academia and industry and also incorporate the need for experts who can not only generate and curate their own data to international standards but also develop software, build new platforms, access high-end computing solutions, collaborate on international projects, manage complex projects requiring diverse teams, and interact with clinicians and public health experts with a clear understanding of patient needs and expectations.

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?

We would highlight the bioinformatics opportunity that exists within 6.3.3 of the Issues paper. While there is a clear vanguard of activity in the medical, agri-bio and health areas, we support the call for more ***integrated approaches to data driven research infrastructure for plant and animal sciences***

*Such infrastructure could support the development of approaches that play a key role in addressing barriers to agricultural yields, including **adaptation and mitigation of climate change**. Just as in the medical and human health area, there will be a greater **integration of genomic, phenomic, metabolomic and proteomic facilities** to capture new scientific advances at the intersection of these areas.*

*Of particular relevance **Agriculture** is a priority given the requirement to maintain food security and the export value as well as the negative impact climate change will have on agriculture. Investment from genomics and phenomics through to precision farming and optimising land use would lead to tangible benefits.*

Australia requires a coordinated infrastructure that supports our national interests in this critical area.

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

***ELIXIR is a broad based life sciences data infrastructure** that focuses on the range of application areas including **health, the environment and natural resources**. There are also various international initiatives in **crop genomics for food security** which Australia should be making significant contributions towards.*

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?

No response

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?

No response

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?

We would highlight the bioinformatics opportunity that exists within 9.1.1 and 9.3.1 of the Issues paper.

*Bioinformatics is of increasing importance in order to **enable the timely evaluation of potential threat that new strains** for example may pose to food safety and biosecurity by rapidly delivering a complete genome, finished genome sequence. Such process and actual data is only valuable as the bioinformatics required for annotating the genome is at hand. Such process involves looking at the genomic sequence, and working out where the genes themselves are, and what they may do. This is done by comparison with genes of known function (for example, from other bacteria).*

*Similarly for **disease control**, bioinformatics is being used to adapt miniaturised sequencing device to conduct **live environmental surveillance**, which enables researchers to deliver real-time experimental genetic data for immediate analysis. Such bioinformatics applications can also be used to develop sensing system by sequencing environmental samples, containing DNA from hundreds or thousands of different organisms.*

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?

*While this section address the fundamental underpinning infrastructure currently delivered through projects such as ANDS, RDS and NeCTAR, it is important to recognise that some of the areas of life sciences have particular privacy and access requirements that need to be supported with the Underpinning Research Infrastructure. Ideally it would be of benefit for this Underpinning Infrastructure to provide solutions that will facilitate and expedite the increasingly **complex***

landscape of data and deliver the dual purpose of protecting and promoting the use and reuse of data*.*

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

As described above, ELIXIR provides a ***framework across the research stack, including some of the Underpinning Research Infrastructure layers****. Key outcomes of the investments of Australia into participating into the international research infrastructure such as ELIXIR would be:*

1. *establishment of an Australian services platform that enables first-class collaboration and partnership with world-leading data infrastructures*
2. *forging the best ways to engage with these rapidly moving data infrastructures to provide value to Australian researchers*
3. *identifying the requirements of Australian researchers for access to world's best practice method and data services in the life sciences and implementing these solutions*
4. *an understanding of the ways in which formal relationships with ELIXIR and BD2K will benefit Australian industry.*

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

No response

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?

Increasingly, the application of biological methods and knowledge drawn from the biosciences becomes of extensively economic relevance. Bioinformatics will underpin these developments.

To support this ***Australia needs a bioinformatics infrastructure*** *to enable research as well as optimal and efficient usage of the discoveries and findings to meet the bio-economic demands. To support the national approach, the bioinformatics research environment must deliver the development and implementation of flexible pipelines which run harmoniously with data management, analysis and statistical tools and methods. It is therefore* ***necessary to establish a bioinformatics infrastructure that is federated to local nodes that act as expert-centred, well equipped and specialised and together worked in a cooperative manner through a coordinated hub.***

This bioinformatics infrastructure would address the following needs of:

1. *networking and funding local centres of expertise for the purposes of ensuring the development of technology*
2. *increasing knowledge transfer between biology research and bioinformatics*
3. *establishing standards for producing, analysing and storing data*
4. *making the necessary software tools freely available and standardising interfaces*

5. *development of long-term strategies for research, action, and funding, in order to improve the conditions for joint public-private funding of collaborative projects*

6. *promote the sustainability of available data resources*

7. *optimisation of the use of computing capacities, in order to improve the utilisation of local resources through comprehensive resource mapping*

8. *contributing towards the improvements in the conditions for transferring data via cloud computing.*

In Australia, the current bioinformatics infrastructure is insufficient for meeting the needs of research, and is probably a substantial limiting factor for the optimal future use of the entire bio-economic potential of modern biosciences. The various fields of biology research – from basic research to applied research – show a similar need for action in the area of bioinformatics.

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

Bioinformatics today, is at the core of life sciences, agri-food and medical research *and will increasingly play a* ***major role in environmental and even biosecurity applications****. The era of data driven research has witnessed the emergence of international efforts, designed to equip their research communities with the appropriate research infrastructure.*

This is the ***perfect time and opportunity for Australia*** *to become a partner in these global initiatives. We can use these initiatives as the framework for establishing a best practice infrastructure within Australia, maximising the investment in the infrastructure from an Australian Government perspective to deliver on the national and international objectives.*

As a formal and recognised member of a program such as ELIXIR our national infrastructure would extend beyond existing infrastructure strategies to provide a distributed infrastructure for life-science information by working together with the European partners:

1. *in the development and adoption of a transparent route to tools and services for data access and exploitation by users through a national discovery portal*

2. *establishing the technical infrastructure across Australia to enable effective data deposition, access, exchange and compute*

3. *establishing the common Interoperability backbone between Australia and ELIXIR encompassing standards, and services that implement the standards, for data archiving, integration and reuse, and*

4. *stimulate innovation and knowledge exchange through training and knowledge exchange.*

Such investment would see Australia benefit from ***first-class participation*** *in, and* ***access to, world-leading data infrastructure for the life sciences.*** *At national level, it would translate into* ***best practice methods and data services available to researchers nationwide****. This would bring greater* ***international and national recognition of Australian*** *services, standards and technologies and increase demand for international collaborations and opportunities for funding.*

Although some efforts in this exists at national level in Australia, there is currently a clear ***lack of awareness and applicability across the life sciences community*** *(when compared to international efforts such as Data Commons (e.g. https://gdc-portal.nci.nih.gov/), http://www.dtls.nl/fair-data/, http://www.dcc.ac.uk/).*

Generally, the provision and use of common and unambiguous identifiers for bio-molecules such as genes, proteins, metabolites and bioactive compounds is key to supporting the information flow from basic science, model organism biology, bioinformatics and structural biology through to translational research and clinical care. There are several international research infrastructures that have been working in this area such as BioMedBridges http://www.biomedbridges.eu/ which has as its mission, to address through a wide variety of initiatives and resources, the ***harmonisation of existing standards and identifiers and the identification of new identifiers where needed, the provision of a registry for users to be able to find suitable standards, and a registry of tools and data resources.*** *Such effort is linked within ELIXIR and of vital importance in forging longer-lasting important collaborations and contributing to the* ***harmonisation of the life sciences data interoperability landscape in which the Australian participation was missing****.*

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?

One aspect that is of concern is the data architecture that is shown in the figure on Page 48. While we accept that all of the key components have been identified, the way that the figure is presented would lead one to believe that all access and collaboration will need to go through a trusted data centre.

We accept that the principles or metadata, preservation, curation, provenance, and authentication are fundamental to the creation of a data infrastructure - we do not believe that this would provide the best architecture arrangement for the next generation for our national infrastructure and fear that it would not provide the best incentives and support for the collaboration and sharing of best practice across capabilities and within research domains. The proposed federated model is designed to implement such standards through contracted projects containing national deliverables.

Other comments

Life Sciences research increasingly relies on access to global data and tools.

Australia's many thousands of life sciences researchers now operate in a global data environment which is increasingly dependent on major pan-national life sciences data infrastructure initiatives: European Life Sciences Infrastructure for Biological Information (ELIXIR: https://www.elixir-europe.org/), Big Data to Knowledge infrastructure (BD2K: https://datascience.nih.gov/bd2k) and CyVerse (http://www.cyverse.org/about). These world-leading data infrastructures have developed in response to critical demand by researchers across Europe and the USA for access to the world's public biological and biomedical information, and the tools, standards and expertise necessary to effectively use it across biomedical, scientific and agricultural research and industry.

Full list of node Partners represented in this submission:

EMBL-ABR is structured as a hub/nodes model, with the EMBL-ABR Hub hosted at the Victorian Life Sciences Computation Initiative (VLSCI), University of Melbourne. Director Andrew Lonie, Deputy Director Vicky Schneider.

EMBL-ABR: ANU Node, Head of Node: Sylvain Foret

EMBL-ABR: UWA Node, Head of Node: David Edwards

EMBL-ABR: VLSCI Node, Head of Node: Andrew Lonie

EMBL-ABR: UTAS Node, Head of Node: Jac Charlesworth

EMBL-ABR: Monash Node, Head of Node: Steven Androulakis

EMBL-ABR: JCU Node, Head of Node: Ira Cooke

EMBL-ABR: AGRF Node, Head of Node: Sonika Tyagi

EMBL-ABR: SBI (UNSW) node, Head of Node: Marc Wilkins

EMBL-ABR: QCIF node, Head of Node: Rob Cook

EMBL-ABR: MA node, Head of Node: Malcolm McConville