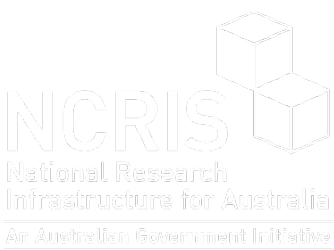
Australian Research Data Commons

Cross-NCRIS EOI Submissions

Table of Contents



Project Title: Australian Urban Health Indicators (AUHI)

Contact Information:

Public contact information has not been provided.

Lead NCRIS facility:

AURIN

Partner NCRIS Facilities:

PHRN

Collaborators:

- Centre for Big Data Research in Health (UNSW), Advisory Committee Participant
- ARC Centre of Excellence for Mathematical & Statistical Frontiers, Advisory Committee Participant
- Melbourne School of Population and Global Health, Advisory Committee Participant
- Digital Health CRC, Advisory Committee Participant

New Data Opportunities:

This project will develop new indicators of health and its determinants by integrating health, socio-economic and other urban data sets to provide a more holistic spatially-explicit understanding of the mental and physical health of the urban population. Geocoded health data held by PHRN participants, such as hospital admission records, will be aggregated and spatially generalised, to ensure individual confidentially, and analytically combined with AURIN spatial data-sets on population demographics, employment (e.g., by sector), economics (e.g., home ownership, rental, household income) and social infrastructure (e.g., access to services, density, housing quality). By systematically integrating data on the social determinants of health with traditional health service planning data, new indicator data-sets will be developed. Examples include indicators that capture (i) vulnerability to contagious diseases, (ii) environmental and physical drivers of long-term health issues, and (iii) the compounding effect of the built environment on mental health and well-being. Such new indicators will allow health and social-service planners to develop spatially targeted policies.

Project Description:

Indicator selection: Existing and new stakeholders, in research, government, health & social services will define key indicators, and scope pilot case studies, to guide selection of data and methods of integration.

This will draw on expertise of PHRN & AURIN's research communities to form an advisory committee.

Ethical & legal issues: We will ensure that, when integrated, the unit-record data held by PHRN cannot be re-identified and develop user policies that recognise the privacy, confidentiality and sensitivity of the datasets and the risk of misuse of the new indicators. PHRN will contribute existing expertise in health data ethics to overcome this challenge.

Indicator development: AUHI will develop programmatic tools that aggregate and integrate the data to generate and publish geo-coded indicator values, while ensuring security and confidentiality. AURIN will contribute existing expertise in spatial data analytics.

Case studies: Generically transferable methods developed will be applied to at least 2 case studies defined in stage 1. The utility of the developed indicators will be rigorously evaluated, drawing on expertise within PHRN's & AURIN's research communities.

Research outcomes:

The benefits of integrating AURIN and PHRN data are plentiful and powerful;

AURIN currently provides data on: accessibility to medical and social services; infrastructure that facilitates wellbeing, such as open green spaces and sports facilities; and factors which might negatively impact health and wellbeing, such as exposure to heavy industry and high-volume road networks. PHRN offers researchers access to individual (unit-record) level characteristics and health outcomes. Integration of the datasets held in AURIN and PHRN will enable a more rigorous quantifiable analysis of the environmental and urban design factors which are determinants of health and wellbeing.

As an example, using AURIN data, researchers from the University of Queensland, the University of Melbourne, and the Royal Melbourne Hospital recently investigated the spatial distribution of child care centres in relation to traffic volumes across Melbourne. It is hypothesised that exposure to vehicle emissions increases rates of childhood asthma. Integrating PHRN data would allow researchers to test this hypothesis fully; accessing hospital admissions for asthma in the children attending those centres.

The AUHI project will deliver:

- An enhanced and streamlined ability for researchers to investigate issues such as determinants of health and wellbeing and health vulnerability in Australia's towns and cities, through seamless, secure and reliable integration and aggregation of AURIN and PHRN data;
- New derived indicator datasets that capture regional propensity to diseases based on environmental and built-environment determinants;
- Novel research case studies and publications.

Areas of research focus (i.e. research questions addressed by the project) could include:

- The impact of disaster events (such as bushfire, drought, pandemics) on mental health.
- The impact of accessibility to open spaces and sports facilities on obesity and diabetes.
- The role of transportation networks in the spread of influenza and other contagions.
- The impact of overcrowding on vulnerability to contagions and pandemic disease.
- The impact of exposure to heavy traffic or industry on diseases such as asthma or cancer.

The AUHI project's new indicators will provide a unique detailed view of the vulnerability of Australia's towns and cities to these impacts. The work will be methodologically novel, allowing applied researchers to access and gain new insight into the spatial characteristics of Australia's urban health. AURIN and PHRN will jointly publish the methodological development of the work within their respective domains. It is envisaged that downstream applied research publications from academics utilising the new data-sets will be substantive within public health, social sciences and the built environment domains. The outcomes will also

enable more-informed evidence-based planning and policy making of health and medical services in Australia.

The AUHI project will act as 'launch pad' for future research grant applications that include (i) scaling of pilot studies to the whole of Australia via focused state funding applications, (ii) technical expansion with the development of a spatially explicit decision support platform for the Australian Urban Health Indicators, and (iii) Applied projects developed by Australian academics from the fields of health studies, social sciences and the built environment.

Broader impact:

The AUHI project will deliver indicators on the health of the urban populous that extends beyond the benefits of any individual research effort; developing health and well-being indicators that will support health service planning and the mitigation of future urban health vulnerabilities. For example, by integrating data from PHRN, such as data on contagious diseases, with AURIN data on overcrowding, availability of jobs and housing, population change and migration, and the location of health services and hospitals, improved indicators of population vulnerability to future contagious diseases across Australia's towns and cities can be developed in order to recognise potential 'hot spots' of vulnerability that may overwhelm local healthcare systems.

FAIRness:

The integrated data (including the derived Indicator datasets) will be expertly aggregated and curated with application of FAIR principles wherever possible.

Project Length:

Two years

Funding Request:

$300 - 400k

Value:

$775,000

Further information:

Other potential partners

- State Health Departments
- Australian Institute of Health and Welfare (AIHW)

Australian Research Data Commons

Project Title: Enhanced data assets for Genomic Medicine - Integrating clinical and experimental genotype-phenotype data for biomedical discovery and disease management

Contact Information:

Michael Dobbie michael.dobbie@anu.edu.au

Lead NCRIS facility:

Phenomics Australia (formerly Australian Phenomics Network)

Partner NCRIS Facilities:

Phenomics Australia

Collaborators:

- Public Health Genomics Program, Monash School of Public Health and Preventative Medicine, Monash University|Lead for the human clinical cohort (ASPREE) dataset
- Genome Informatics Group, John Curtin School of Medical Research, Australian National University|Bioinformatics lead for the experimental (mouse) dataset

New Data Opportunities:

- Overview – Complementary data assets will be linked/integrated for managed access through a portal to create a new and informative national data assets.

- Context – Australia is embarking on its Genomic Health Futures Mission ($500M over 10 years from the Medical Research Future Fund), as prioritised in Australia 2030: Prosperity Through Innovation (ISA, 2017) and The Future of Precision Medicine in Australia (ACOLA, 2018). Phenomics Australia will provide the necessary NRI.

- Challenge – Improved functional annotation of clinical genomic datasets is required to address the emerging twin challenges the exponentially increasing number of human variants of unknown significance, concomitant with insufficient specific disease biomarkers.

- Solution – Integration of clinical and experimental organism genotype-phenotype datasets will

better inform clinical decision making (for disease diagnosis and treatment) and experimental design (for fundamental biological discovery of the genetic determinants of health and disease and biological insight for the discovery of new drug targets). Data infrastructure (high quality data and a linkage platform) is key.

Project Description:

The project will develop and implement the framework and platform for linking/integrating two NCRIS-enabled, generated and/or managed datasets:

1. Human clinical dataset: ASPREE-G (Genomics) and the associated Medical Genome Reference Bank dataset that consist of matched genomic and phenotypic data, which resulted from a bi-national clinical trial and longitudinal study of healthy aging. Major co-investors include BPA.
2. Experimental mouse datasets: Missense Mutation (mouse SNP) Library and Australian Phenome Bank, both Phenomics Australia-enabled, generated and curated.

Working with a team of bioinformaticians/data scientists the project will:

1. Acquire/develop data management resources/tools to ready data for automated analysis (such as by ML protocols), including minting DOIs and metadata standards;
2. Develop a data portal to enable data FAIRness, interfacing clinical genetics consortia, and ensure visibility/integration across national and international boundaries; and
3. Develop impact measures.

The resulting data infrastructure platform (data asset) will be sustainably governed and scalable, designed to maximise integration of data sets with wider genetics datasets.

Research outcomes:

A deeper functional understanding of the genomic contributions to health and disease will synergistically boost clinical management and biomedical research, creating formal linkages between model organism research communities and databases with Genomic Medicine initiatives leading to higher impact publications and better designed research projects.

Broader impact:

Improved diagnostic decision making, targeted therapeutic development, and design of new disease prevention strategies.

FAIRness:

Human and experimental datasets will be made FAIRer and will support computable phenotypes for improved reproducibility and reliability.

Project Length:

Two years

Funding Request:
$300 - 400k

Value:
$4,000,000

Further information:
Regarding Project Value: All other existing NCRIS investments = >$4M (cash and in-kind), including $750K cash contribution from BPA for generation of ASPREE SNP dataset, >$3M cash contribution from Phenomics Australia for the generation and curation of the mouse datasets and operation of the Australian Phenome Bank. Additionally, in-kind contributions from appropriately-skilled staff at ANU and Monash to support this project.

Additional information provided upon request from ARDC:
Phenomics Australia will partner with Bioplatforms Australia (BPA) to achieve the aims of the project.

Together with BPA and the custodians/curators of the ASPREE-G clinical dataset, Monash University/UNSW, we will (1) assess the content and format of the clinical and experimental datasets to plan and implement data linkage (including considerations of metadata standards, DOI requirements, enabling automated analysis, and scaling to incorporate additional data), (2) investigate the characteristics of a data portal, as well (3) develop impact measures and a sustainable governance framework.

BPA will be providing the clinical genotype data in support of the ASPREE phenotype data to form the ASPREE-G dataset that is part of the Medical Genome Reference Bank dataset, which consists of genomic (common gene variants / human Single Nucleotide Variant (SNPs)) analysis matched to clinical phenotype data, which resulted from the bi-national ASPrin in Reducing Events in the Elderly (ASPREE) clinical trial and longitudinal study of healthy aging. In partnership with Monash University and the Ramaciotti Centre (UNSW), BPA contributed $750,000 towards the cost of DNA sequencing of the ASPREE samples to generate the ASPREE-G SNP dataset.

Together, BPA and Phenomics Australia will partner and offer substantial in-kind co-investments to enable this project. These key investments are the expertise of our staff, our partner network, and researchers that will be required to oversight the work packages, assess the existing data, conduct the project, and test the market, ensuring the FAIR data principles have been achieved.

Project Title: Disparate Biomedical Data Assets

Contact Information:

Graham Galloway g.galloway@uq.edu.au

Lead NCRIS facility:

National Imaging Facility

Partner NCRIS Facilities:

Bioplatforms Australia
National Computational Infrastructure

Collaborators:

- Australian Cardiovascular Alliance|Work Package Lead for Cardiovascular dataset
- Austrlian Centre of Excellence in Melanoma Imaging and Diagnosis|Work Package Lead for Dermatology Data
- The Florey Institute of Neuroscience and Mental Health|Work Package Lead for Epilepsy Data
- Australian Brain Alliance|Work Package for Brain Data

New Data Opportunities:

This project will integrate phenomic and genetic data with clinical and pre clinical correlates from Australian communities committed to the FAIR data principles. Each of the collaborators will deliver major data assets from distributed projects, such as neuro, cardiovascular and dermatological normative data, as well as data from key clinical cohorts. Data collected by the project will be findable by other communities who have similar needs for aggregated data patient/subject/sample across domains.

The data assets provided by the partners are:

Australian Cardiovascular Alliance - leading a 10 MRFF year research program to find solutions to cardiac disease.

Australian CoE in Melanoma Imaging and Diagnosis - ACRF project to build a database of 360° images of human dermatomes, with longitudinal data for patients susceptible to skin cancer.

Australian Epilepsy Project - a MRFF project planning to build a national database of multivariate brain images, genetics and cognitive testing in people with epilepsy.

Australian Brain Alliance - national resource of preclinical and human brain microscopy, in vivo images and electrophysiology data, correlated to genomic and clinical data.

Project Description:

Precision medicine and big data are the buzz-words of 21st century medical research. Many research groups developing AI platforms recognise that they cannot, alone, acquire sufficient data to train and test. Nor does the data have the diversity necessary to build clinical products. A crucial need for the future of medicine is aggregated data assets from many technologies and multiple types. The project data assets include imaging, genetic, clinical, and environmental data. As a coordinated project, the scope is to define clear access mechanisms to address ownership, while facilitating collaboration, sharing, and sensitive data management. Establishing mechanisms to describe data using community agreed metadata standards, to ensure interoperability. The key deliverable is a compendium of databases that are accessible through interlinked national repositories. To ensure requirements of ethics in different jurisdictions are attended, this needs to be federated, thus allowing data to be shared outside of their jurisdiction. In the later stage, we aim to make project datasets available in appropriate computational technologies to facilitate processing.

Research outcomes:

The four collaborators are research organisations responding to competitive grant schemes, in which publications are a top priority to maximise the value of research funding. Federating efforts to interlink data assets, currently residing in the literature, will enable the scientific exploitation of phenome data via accessible repositories. This will result in new projects to set data sharing base lines across institutions, resulting in broader opportunities for collaborative research projects resulting in publications and grants.

Broader impact:

We aim to offer a comprehensive, publicly accessible compendium of genes and genetic phenotypes related to clinical and preclinical studies. Datasets will be accessible to registered users, and the community will be able to filter for gene-phenotype relationships based on their research scope or their interest to collaborate. In addition, defining inherent relationships and hierarchies between collaborator data assets is fundamental to enable clinical data analytics, decision support, and artificial intelligence for health and societal impact.

FAIRness:

All stakeholders are committed to making data available to the wider scientific community, national and international. Effective text mining is needed to gather these data assets. A prerequisite is the availability of specified domain vocabularies and catalogues that are shared across the different data generators. We will build upon existing community guidelines, such as the NIF Trusted Data Repositories (ANDS/NeCTAR, 2017), to ensure reliability of data, including associated QC data, and rich metadata to describe the acquisition conditions and the clinical state of the participants.

Project Length:

Two years

Funding Request:

$300 - 400k

Value:

$800,000

Further information:

Whilst the platform will be used to create separate repositories for different domains/stakeholders, Australian Brain Alliance, Australian Cardiovascular Alliance, Australian Centre of Excellence in Melanoma Diagnosis and Imaging, it is envisaged that there would be the ability to federate across instances, requiring rich metadata including informed consent.

The platform would not be restricted to biomedical and would support biodiversity and agriculture repositories, both of which need genomic/phenomic connectivity.

Additional information provided upon request from ARDC:

The aim of the project is to link datasets being acquired across multiple NCRIS capabilities. BPA and NCI will provide access to genetics, metabolomics and proteomics data, including the workflows for curating that data. The project will ensure data is curated with rich meta data, such that it can be found and shared. Importantly, this project will ensure that the data is exposed to suitable compute infrastructure for analysis and machine learning. So draft of work packages:

WP1 – Curation of Imaging data – NIF – integration of Trusted Data Repositories and Australian Imaging Service

WP2 – Curation of Genetic data – BPA and NCI – Integration of BioCommons

WP3 – Curation of associated Clinical data and association with imaging and genetic data – NIF, BPA – integration with partner projects: ACvA, ACEMID, AEP, ABA

WP4 – Exposure of data assets to compute infrastructure and workflows – NIF, BPA, NCI

As such, the capabilities do not provide the data. As an infrastructure provider, the data belongs to our users. This project will deliver the infrastructure to allow those users to collaborate across disciplines. The partner projects are all multisite, multi-modality projects which have identified the need for systems to manage disparate data types. This is something that is currently developed at an individual project level, leading to duplication of effort, which develop bespoke systems with different data standards and storage technologies. The proposed project will

a) Deliver integrated data management and curation for the partner projects, and exposing this data available to the wider research community

b) Establish a framework for a more general system of managing data from different sources

Australian Research Data Commons

Project Title: Biodiversity data integration to support national environmental reporting (State of the Environment)

Contact Information:

Andre Zerger ala@csiro.au

Lead NCRIS facility:

Atlas of Living Australia

Partner NCRIS Facilities:

Integrating Marine Observing System
Terrerstrial Ecosystem Research Network

Collaborators:

- Department of Agriculture, Water and Environment (DAWE)|Business owner & policy lead
- Griffith University - EcoCommons Program|Computational and data delivery partner

New Data Opportunities:

DAWE produces the State of the Environment (SoE) report every 5 years for statutory reporting obligations and to update all Australians and decision-makers on environmental state, pressures, trends and key issues. SoE relies extensively on high-quality national data. Given the increasingly rich data from NCRIS facilities since the 2016 report, SoE is a perfect use case to develop new cross-facility data assets to support national environmental reporting. Integrated data products will have significant value beyond SoE with use in the research sector, in related government programs and for international reporting obligations (e.g. UN SDGs, CBD Aichi Targets).

ALA, TERN and IMOS serve biodiversity and environmental data and are conduits for data use. SoE reporting will build data connectivity to develop policy- and industry-ready datasets to support contemporary reporting and analysis needs. The project will address data management policy for use of sensitive data in products, analysis and reporting. The project could be extended in future to accommodate relevant data from other NCRIS facilities, notably BioPlatforms Australia, AURIN and site-based data from the agricultural sector.

Project Description:

The cross-cutting biodiversity needs of the 2021 SoE will serve as a focus for new data products supporting national reporting and to prioritise data linkages and summary formats. Resulting data pipelines will be maintained for future continuous integration of data from partners. Future SoE reports will be digital and able to offer consistent repeated summaries. Pipelines will be generalised so researchers can extract data in the same format for an arbitrary time period and land unit.

Funds will be used for:

* Data requirements/formats
* Pipelines and workflows to generate datasets with metadata on provenance and known issues
* Validation and integration into SoE

Stage 1: SoE requirements and planning

* Evaluate data requirements for SoE working closely with biodiversity chapter author
* Review relevant partner data assets
* Design data products, policy and APIs

Stage 2: Pipeline development

* Prepare data streams for integration
* Develop data processing workflows
* Establish continuous generation of products
* Validate data product via early access

Stage 3: Evaluation of data products

* Ensure data products meet DAWE requirements
* Report developed data products

Research outcomes:

Beyond SoE, cross-facility data assets will be usable in biogeography and ecology research, biodiversity assessments and conservation planning.

This project will demonstrate organisation of complex data from multiple sources to support environmental reporting.

The project is relevant to international programs. For example, the Global Biodiversity Information Facility aspire to produce report-ready data products, so this work is anticipated to have global impact. The process and products will therefore be documented in a suitable international journal publication.

Broader impact:

This project focuses on the need for Australian environmental and biodiversity research infrastructure to deliver data to support the SoE and similar reporting by State and Territory governments. However, the data products are likely to deliver additional benefits for related activities such as DAWE's environmental-economic accounting program and ABS land and ecosystem accounts. Incorporating robust nationally relevant biodiversity data is a key data gap in these related programs so the development of such a dataset is likely to deliver significant impact.

FAIRness:

Data products will receive DOIs and documented with metadata conforming to standards used by the partner NRIs, including provenance of the dataset.

Data and metadata will be freely accessible for download from partner portals.

Data products will use data standards applied by ALA, TERN and IMOS.

Data products will be publicly accessible and licensed under the most open Creative Commons licence applicable. Some potentially relevant datasets are licensed as CC BY-NC, but most data is CC BY or CC0. Provenance metadata will document data origins and transformations.

Project Length:

One year

Funding Request:

$300 - 400k

Value:

$800,000

Further information:

This work will extend linkages to other analytical and reporting streams including with the national State of the Environment Reporting Forum with State and Territories.

This work will augment multi-disciplinary environmental information and reporting and its contribution to our understanding of environmental health for human health and resilience, particularly given the recent examples of bushfire events.

This work is timely and aligns with global objectives for the Strategic Plan for Biodiversity and the post-2020 framework.

Australian Research Data Commons

Project Title: Data nexus: coupling genomic and environmental data to enhance integration

Contact Information:

Michelle Heupel michelle.heupel@utas.edu.au

Lead NCRIS facility:

IMOS - UTAS OFFICE

Partner NCRIS Facilities:

Integrated Marine Observing System (IMOS)
Bioplatforms Australia (BPA)

Collaborators:

- University of Technology Sydney|Development lead
- University of Queensland|Contributors to requirements
- CSIRO|Contributors to requirements

New Data Opportunities:

Microbes are fundamental to human and ecosystem health. Our ability to understand and predict the response and resilience of marine microbial communities is essential to conserving a healthy marine environment, and the subsequent social, health and economic benefits we receive.

Data resources have been developed across Australia that capture molecular microbiological and genomics data (supported by Bioplatforms Australia) as well as detailed oceanographic data (supported by IMOS). The challenge remains, however, to bring these data together in a standardised and interoperable manner.

Coupling genomic and oceanographic data in a form that is highly usable and intuitive for non-experts will greatly enhance integration between disciplines, accelerate discovery for researchers and render a large dataset into a format that is engaging and accessible for diverse users. The new data asset will reveal the intricately coupled physical and biotic mechanisms that underpin ocean health and productivity and help infer the status of these critical services under future ocean scenarios.

Project Description:

This project will bring together complementary but disparate datasets to create an enhanced asset for the assessment of microbial communities and their corresponding environment. We will develop an interoperable data management service that builds on existing metadata standards between the Australian

Microbiome Initiative (Bioplatforms Australia), in conjunction with detailed oceanographic data generated from IMOS National Reference Stations (AODN, IMOS). This will be supported by an easily accessible and highly interactive online interface for the coupled visualisation of genomics and molecular microbiological data.

Additionally, the combined data will be processed to produce secondary spatial assets. For example, it will be possible to generate graphical time-series that overlay molecular indices of microbial biodiversity onto physical oceanographic data. Machine learning approaches will enable distribution models of functional capacity to be overlayed onto maps of oceanographic patterns such as current vectors. This will allow users to clearly visualise temporal and spatial patterns in marine microbiological data and directly relate these to oceanographic conditions.

Research outcomes:

Integration and visualisation of the vast ocean data from both NCRIS capabilities will provide a highly novel framework for inferring the health and productivity of current and emerging ocean ecosystems. Such a utility will enable meaningful, cross-disciplinary collaborations between researchers investigating the ecological and socioeconomic impacts of an array of events such as bushfires, marine heatwaves, algal blooms and long-term ocean warming. The framework would also support ground-truthing and validation of the next generation of high-resolution earth system models.

Broader impact:

Developing our understanding of the role of microbial communities, and how this role may change under differing physical, biological and environmental conditions will generate information that will be meaningful across a number of industries including health and wellbeing, tourism, fisheries and agriculture, conservation, mining, as well as disaster recovery and remediation (e.g. oil spills). The project team will aim to work with researchers across these groups to facilitate the use of the data asset.

FAIRness:

The proposed data asset will be:

- Findable through well described, searchable metadata and an interactive user-friendly interface;
- Accessible through open access machine-to-machine web-protocols;
- Interoperable by using agreed formats and standardised metadata based on community best-practice; and
- Reusable through suitable data licencing and contextual information so users can ascertain if the data is fit for their purpose.

Importantly, this project is not set to replace or reinvent existing repositories but rather leverage them through an aggregation layer with harmonised metadata

Project Length:

Two years

Funding Request:

$300 - 400k

Value:

$800,000

Further information:

It is notable that in many respects microbiological datasets, such as species and gene abundances tables, are very similar in format to other biological observational data, but differ in their breadth and complexity (thousands of species, millions of genes) and therefore have potential to be analysed and modelled with common ecological statistical approaches (such as machine learning) and/or to leverage existing computational approaches and infrastructure. We anticipate that this resource will generate new information allowing the resource to continually grow.

Australian Research Data Commons

Project Title: OzBarley: from Genome to Phenome and back again. A barley data and germplasm asset for the Australian research and breeding community.

Contact Information:

Bettina Berger bettina.berger@adelaide.edu.au

Lead NCRIS facility:

Australian Plant Phenomics Facility

Partner NCRIS Facilities:

Australian Plant Phenomics Facility (APPF)
Bioplatforms Australia (BPA)

Collaborators:

- CSIRO|panel design & curation, data science strategy, germplasm stewardship system, pedigree visualisation
- University of Adelaide|panel design & curation, data science strategy
- Australian Grain Technologies (AGT)|single seed decent cleaning of panel
- Intergrain|supporting field phenotyping
- South Australian Research and Development Institute (SARDI)|seed bulking and field trials
- Australian Grains Genebank (AGG)|seed distribution to community
- Federation University|data stewardship, link to AgReFed
- Australian Genome Research Facility (AGRF)|genotyping, sequencing

New Data Opportunities:

Building greater resilience in crops is critical to provide food and agricultural products in a sustainable manner for a growing human population. This is particularly significant as agriculture is facing increasing challenges with a changing climate. Breeders will often select for individuals with the highest yield potential, however certain genotypes may perform well in one environment but poorly in another. Thus, correlating the genotype with diverse phenotypes helps us to understand and predict how the two interact in different field conditions, and ultimately how this will impact productivity.

Using barley as an exemplar use case, this project will bring together genotypic and phenotypic datasets made available through BPA and the APPF in close collaboration with research and industry experts. Bringing

this data together will allow researchers, breeders, bioinformaticians and machine learning experts to work together and push the limits of current analysis approaches. By integrating phenotypic with genotypic data we can identify genetic loci for desired crop characteristics and trace them through the breeding pipeline to ultimately deliver improved varieties to growers.

Project Description:

This project will bring together complementary but disparate datasets to create an enhanced data asset for the assessment of crop genomes, and the corresponding variability in trait expression with relation to the environment. Initially, OzBarley will select and curate two diverse barley panels. One panel will capture the Australian breeding history and will draw on and assemble historic pedigree data. A second panel will include landraces containing genetic markers and traits that may have been lost during the breeding process.

The proposed initiative will provide interoperable data management for genome-to-phenome (G2P) assets, enabling users to rapidly identify and link traits with genetic material for research and breeding. This integrative, multidimensional approach to G2P data management will serve as a benchmark for capturing data where both genomic and phenomic measurements are required by researchers.

Importantly, this project is not intended to replace or reinvent existing repositories but rather leverage them through an aggregation layer with standardised metadata services.

Research outcomes:

Generation of the OzBarley panel and associated datasets is in itself a publication opportunity. Future publications analysing the data assets created will span diverse disciplines, such as quantitative genetics, crop physiology, machine learning, computer vision, bioinformatics and statistics. Similar projects (e.g. OzWheat) have been instrumental in Grains Research and Development Corporation tenders being put forward and in improving predictive crop models. Our project team includes researchers and breeders to facilitate the early uptake of these data assets.

Broader impact:

Barley is worth $2.3B p.a. to the Australian economy, with Australia being the second largest exporter. The recent drought has emphasised that in the face of climate change, we rely on resilience and sustainability in the farming sector, and elite crop varieties play a vital part in this. OzBarley will support transformative research through scientific outputs (e.g. gene discovery) and industry applications (e.g. increased rate of genetic gain in breeding). The collaboration between NCRIS, researchers and breeders in OzBarley will enable translation of data into research outcomes and impact.

FAIRness:

OzBarley will use publicly available barley lines to ensure outputs can be shared, and breeders IP will not be affected.

Data will be:

- Findable in a discipline-specific interface, e.g. GERMINATE, meaningful to breeders and researchers;
- Accessible through open access machine-to-machine web-services;
- Interoperable by using discipline standard (meta)data formats and vocabularies; and
- Reusable through open data licencing and contextual information so users can ascertain if data is fit for purpose.

The project will engage with AgReFed to implement appropriate data stewardship principles.

Project Length:

Two years

Funding Request:

$300 - 400k

Value:

$1,000,000

Further information:

OzBarley is an NCRIS enabled project, driven by the research and breeding community, as illustrated in the list of collaborators. While NCRIS will facilitate the generation, curation and sharing of these valuable data assets, the project is shaped through and strongly supported by the domain experts from the public and private sector. This demonstrates the user demand and will ensure uptake of the data assets generated.

Australian Research Data Commons

Project Title: A National Scale Data Asset to Integrate Molecular Imaging with Bio-analytics

Contact Information:

James Whisstock James.Whisstock@monash.edu

Lead NCRIS facility:

EMBL Australia / Monash University

Partner NCRIS Facilities:

Microscopy Australia
Bioplatforms Australia

Collaborators:

- Monash University|Lead, Electron Microscopy
- Monash University|MASSIVE Platform
- Monash University|Lead, Monash Proteomics

New Data Opportunities:

We propose to enable a new, publicly accessible national scale data asset to underpin the integration of molecular imaging with bio-analytics, thus driving discovery research across the whole of the life sciences. The resource will permit Australian researchers to attain one of the great ambitions of biologists and understand the precise molecular make-up of the intracellular milieu. We will use Artificial Intelligence (AI)-driven bioinformatics approaches to seamlessly integrate and interrogate high-resolution imaging data (derived from optical and electron microscopy (EM) and X-ray crystallography) with proteomic/genomic data and gene ontology/protein interaction network data. Currently, this specialized information is distributed across numerous disparate databases, precluding the ready interpretation and analysis of imaging data such as 3D tomograms output by the latest generation optical and electron microscopes. Our online platform will host final, released annotated datasets and permit presentation of the data to the community. Our work will have immediate application in fields such as drug discovery, the study of infectious disease and in molecular diagnostics.

Project Description:

EM has advanced to the point where it is possible to determine the 3D structure of individual proteins in situ (i.e. in the context of parts of intact cryo-preserved cells). Currently, however, a key limitation of this

technique is that the identification of proteins in the region of interest is extremely challenging and relies on exhaustive comparative experiments. A potential solution to this problem arises through integration of EM with biological mass spectrometry (proteomics), but this technique is limited to samples far larger than individual cells. Through novel preparation approaches, we can surmount these challenges and gain substantive insight into the proteome of portions of individual cells. The challenge now lies in accurately correlating the proteomic data with the information derived from imaging experiments. We anticipate that this process will be greatly enhanced through the use of AI and via reference to extensive structural biology data and gene ontology/protein interaction network information. Accordingly, we aim to develop a new resource to organize and navigate multidimensional data and drive connectivity between molecular imaging and proteomic datasets.

Research outcomes:

The discovery of the molecular makeup of dynamic networks of biological macromolecules will drive new discoveries across a diversity of life sciences, including in drug discovery, drug target validation, structural biology, agribuisness and diagnostics. Project outcomes will include high profile papers in generalist journals as well the generation of new Intellectual Property. Collaborative exchange across the diversity of the life sciences, as well as deploying new innovations from materials sciences, will underpin new opportunities to win national and international grant funding.

Broader impact:

Combining molecular imaging with bio-analytics promises to revolutionize discovery in biology. The mechanistic insights gleaned through analyses of these data will acceralate and underpin innovation in areas that are important from an environmental, societal and economical perspective. For example, discovery impact on agribusiness will help our nations response to the changing climate and new drugs will impact on the health of society. Collectively innovation more broadly will drive the generation of new jobs and the establishment of new industries.

FAIRness:

FAIRness and Quality

Findable: Data will be published in adherence to a machine readable standard such as bioschemas.org and submitted to a central data record repository (such as ANDS).
Accessible: Monash Uniersity will address requirements for data access, preservation and stability, and will provide a 10 year commitment to host the data.
Interoperable: Data will conform to relevant standards across biosciences, proteomics and imaging (e.g. https://doi.org/10.1038/ng.1054).
Reuseable: Data will be published will relevant provenance information and with an appropriate open license.

Project Length:

Two years

Funding Request:
$300 - 400k

Value:
$1,200,000

Further information:
The proposed research will involve a close collaboration between three NCRIS funded platforms – the Victorian Node of Micrscopy Austalia, the Bioplatforms Australia funded Proteomics facility at Monash University and EMBL Australia (Proteomics, Data Analytics and Bioinformatics). The work will further be closely integrated with the activities of MASSIVE, a computational platform that has a successful and extensive track record of developing data assets and in delivering on ARDC funded projects.

Additional information provided upon request from ARDC:
With regards to Bioplatforms Australia, the proposed contribution they will make to the project is via the equipment and staff they fund within the Monash University Proteomics Unit. This includes technical expertise in regards to proteomics and in respect to the analysis of proteomic data.

The project will use the tools and platforms provided under the ACCS, and it would be challenging to undertake this project without ACCS capability. The ACCS has a work packaged dedicated to big-data producing CryoEM facilities at all flagship high-energy CryoEM installations in Australia: Monash, UQ and UoW.

This work package is focused on mechanics to solve the big data challenges faced by these facilities, including data volume, acquisition speed, data processing requirements, and a complete lack of standardisation in formats and metadata.

This project will co-register data from CryoEM microscopes (using ACCS platforms for capture and analysis) with proteomic/genomic data and gene ontology/protein interaction network data to create a unique data asset.

The project proposal will be provided to the Australian Characterterisation Informatics Committee to ensure strategic national alignment, ensure no duplication and identify opportunities for leverage and collaboration.

Project Title: Building the National High Resolution and High Performance Geophysics Reference Collection for Next Generation Data Analysis

Contact Information:

Ben Evans Ben.Evans@anu.edu.au

Lead NCRIS facility:

NCI

Partner NCRIS Facilities:

Auscope
TERN

Collaborators:

- CSIRO: Users of geophysical data, software and analysis methods. Development of new methods and applications for using high resolution data through the CSIRO Deep Earth Imaging Future Science Platform and the spatiotemporal theme of the Machine Learning/Artificial Intelligence Future Science Platform.
- Geoscience Australia: Provider of data and expertise on data processing
- Universities, Government and Industry Partners of AuScope: Providers of data as well as data processing and new research techniques for the analysis of data at scale
- ARC Linkage Loop project: users of multi-disciplinary data
- Any researchers: wanting to use publicly funded Australian geophysical datasets
- Minerals, Energy, Groundwater industries: Beneficiaries - access to data in rawer forms for more precise targeting
- State and Territory Geological Surveys: users wishing access to high resolution data, as well as potential data providers in their own right
- TERN "both as an NCRIS facility involved in data provider quality (e.g., ASTER) and as end-users."

New Data Opportunities:

Large volumes of geophysical data have been acquired by Universities or Government agencies since 1950. Historically, the raw forms of data (processing levels L0-L2) have been hard to access due to inabilities to manage and analyse data at scale: released data has been highly processed/downscaled, accessible only as

file downloads or visualisations. Some L0-L2 geophysical data has been progressively organised at NCI, with associated HPC/HPD-enabled codes.

The new data opportunities are to:
* Move as many minimally processed, high resolution geophysics datasets to NCI collections, made FAIR, enabled with advanced data protocols, and suitable for scalable computation as a single integrated reference collection of multiple geophysics types (magnetic, gravity, AEM, radiometrics, MT, INSAR, GRACE, passive seismic): apply new methods in ML/DL
* Use identifiers to facilitate better referencing, citation and provenance for embedding within derivative products including data assimilation and/or online data systems of AuScope's partners and Federal & State Government catalogues
* Increase data uptake by curated code and Jupyter Notebooks available for access for in-situ HPC analysis or data services

Project Description:

The project has 3 main components, but not strictly sequential
Data Preparation and Publication
- Priority Datasets identified
- Organise within NCI's data management, including FAIR, provenance, attribution, license
- For each data type enhance/develop a HPD format conformant to analysis-ready standards
- As required, data reformatted into computational self-describing formats
- Datasets QA/QC'ed against benchmark analysis software and standards
- Data Published through NCI Catalogue and ServicesEnd User Community Engagement
- End User consultation of needs that can be brought together around the data
- Release Web area on NCI for user materials and data organisation (similar to https://opus.nci.org.au/display/CMIP/CMIP+Community+Home/)
- Release the curated Jupyter Notebooks (similar to https://nci-data-training.readthedocs.io/en/latest/_notebook/climate/climate_withtable. html), that demonstrate widely relevant analysis cases that demonstrate how to use the capabilities

Wider Data Discovery
- AuScope Portal/AVRE to update with any new data through NCI catalogue
- AuScope partners and government agencies can link to source data for their derivative data products
- NCI catalogue data cross-referenced in ANDS RDA

Research outcomes:

Having ready access to these well-managed data and in a HPC environment will significantly accelerate researchers ability to generate new results and develop and share their techniques.
Machine Learning and Deep Learning Techniques which are currently not possible due to the lack of standards and unification of the data will be enabled, leading to world-leading capabilities, and will be a reference site for other similar activities at overseas infrastructures. The project will liaise with key users to define use cases, such as CSIRO Mineral Resources, the CSIRO Deep Earth Imaging and Machine

Learning/Artificial Intelligence Future Science Platforms, Geoscience Australia and the ANU Research School of Earth Sciences. These examples will also be used to help inform for additional requirements on the data models/structures, and may be included in some benchmark cases.
Advanced data modelling technologies, such as probabilistic inference methods, can be computationally very expensive and will benefit from HPC facilities. Having algorithms and geophysical data hosted on the same HPC environment will ease and speed up the research.
Geophysics data will be more accessible to other disciplines, which are currently unable to engage due to its poor data interoperability and non-analysis ready with more widely used software outside the Earth sciences.
For research teams/the individual researcher, multiple publications should result from increased accessibility of the data: use of PIDs and implementation of FAIR will help researchers meet publisher requirements and help funders/researchers/institutions track impacts from grants.

Broader impact:

Geophysics data has many applications in the Earth and environmental sciences. For example, in the resources sector (minerals, energy, groundwater), geophysics is a critical tool in being able to determine subsurface structures favourable to hosting deposits. Geophysics is a vital tool in minimising societal risk from natural and anthropogenic hazards (landslides, earthquakes, ground subsidence). Combined, these are of economic benefit.
Research teams will be able to analyse large volumes of high-resolution data and see the quality of the algorithms that they are generating very quickly. Rather than using pre-canned workflows, researchers will be able to tune algorithms specifically to the local geological/environmental conditions. A better quantification of the uncertainty and more reliable predictions can be achieved including more parameters during the modelling and using different algorithms.
A longer term broader impact of this project is that by having the higher resolution data adjacent to HPC, then more precise solutions can be made, and particularly for the hazards space, analytics can be done in faster-than-real-time.

FAIRness:

All datasets will comply with the new NCRIS FAIR data guidelines. As both AuScope and NCI signed the Commitment Statement by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS), the project will meet additional requirements of Earth science publishers by each dataset having a PID and landing page.
All data will be validated against the published NCI Data Quality Strategy (https://doi.org/10.3390/informatics4040045), which is deepening as more data examples emerge. The ESIP/ARDC Data Quality CoP resources page already links to NCI data quality strategy for broader use.

Project Length:

Two years

Funding Request:

$300 - 400k

Value:

TBC but anticipate 1:1 pending more detailed planning. Anticipate total value of $700-800K.

Further information:

This builds on major investments led by NCI and AuScope and includes co-investments from NCRIS (including RDSI, ANDS, NeCTAR and ARDC): for almost every data type there is either an operational or prototype model in place. Many of the L0-L2 datasets are already at NCI, but many are in partner storage areas: some are externally held by AuScope and collaborating partners. All data will be made available online through the NCI repository using standardised web services protocols. The data could then be used by other portals such as the AuScope Portal/AVRE, Research Data Australia, National Map.

Additional information provided upon request from ARDC:

https://ardc.edu.au/wp-content/uploads/2020/06/NCI-response-for-geophysics-proposal-to-ARDC.pdf